Comparative Analysis of Deep Learning and Hybrid Models on Multi-Disease OCT Image Classification Datasets with Explainable AI

Dissertation submitted to
Visvesvaraya National Institute of Technology, Nagpur
in partial fulfillment of the requirements for the award of
the degree

Master of Technology
in
Computer Science & Engineering

by

Suvendu Bastia (MT23MCS023)

under the guidance of
Dr. Meera M. Dhabu

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology
Nagpur 440 010 (India)

JUNE 2025

©Visvesvaraya National Institute of Technology (VNIT) 2025

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology,
Nagpur

Declaration

I, **Suvendu Bastia**, hereby declare that this dissertation titled **"Comparative Analysis of Deep Learning and Hybrid Models on Multi-Disease OCT Image Classification Datasets with Explainable AI"** is carried out by me in the Department of Computer Science & Engineering at Visvesvaraya National Institute of Technology, Nagpur. The work is original and has not been submitted earlier, in whole or in part, for the award of any degree at this or any other institution.

Suvendu Bastia
Suvendu.

Name and signature
Date: 16-6-2025

Certificate

This is to certify that the dissertation titled **"Comparative Analysis of Deep Learning and Hybrid Models on Multi-Disease OCT Image Classification Datasets with Explainable AI"** is submitted by **Suvendu Bastia** in partial fulfillment of the requirements for the award of the degree of **Master of Technology in Computer Science & Engineering, VNIT Nagpur**. The work is comprehensive, complete, and fit for final evaluation.

Dr. Meera M. Dhabu
Department of Computer Science & Engineering, VNIT Nagpur

Dr. M. P. Kurhekar
Head, Department of Computer Science & Engineering, VNIT Nagpur
Date:

Department of Computer Science & Engineering
Visvesvaraya National Institute of Technology, Nagpur

Non-Plagiarism Certificate

Certified that M. Tech Dissertation titled " **Comparative Analysis of Deep Learning and Hybrid Models on Multi-Disease OCT Image Classification Datasets with Explainable AI** " submitted by **Suvendu Bastia**, Enrollment No. **MT23MCS023**, ID No. **30745** has been checked for plagiarism using TURNITIN software and the overlap is found within prescribed limits. A summary of the report is given below.

ORIGINALITY REPORT

15%	11%	12%	8%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Suvendu

Suvendu Bastia
Enrollment No: MT23MCS023

Dr. Meera M. Dhabu
Associate Professor
Department of Computer Science & Engineering
VNIT, Nagpur

ACKNOWLEDGEMENT

I would like to express my deepest gratitude and sincere appreciation to all those who have supported and guided me throughout this M.Tech journey.

First and foremost, I am immensely grateful to my supervisor, **Dr. Meera M. Dhabu** Ma'am, for her invaluable guidance, expertise, and encouragement. Her constant support, patience, and insightful feedback have been instrumental in shaping this thesis. I am truly fortunate to have the opportunity to work under her mentorship.

I would like to extend my heartfelt thanks to **Dr. M. P. Kurhekar**, Head of Department, for his invaluable support and guidance throughout this journey. His encouragement has been vital in the successful completion of this project.

I would like to extend my heartfelt thanks to the faculty members of the Department of Computer Science & Engineering at Visvesvaraya National Institute of Technology Nagpur for their knowledge-sharing and for providing a conducive academic environment. I am indebted to them for their teachings and guidance which have broadened my perspective and enhanced my understanding of the subject. My sincere appreciation goes to them for their assistance, collaboration, and providing necessary resources for conducting experiments and carrying out this research work. Their contributions have been significant in the successful completion of this thesis.

I would also like to express my gratitude to my family and friends for their unwavering support, understanding, and encouragement throughout this journey. Their love, motivation, and belief in my abilities have been the driving force behind my accomplishments.

Thank you all.

Suvendu Bastia

Abstract

This study, titled **Comparative Analysis of Deep Learning and Hybrid Models on Multi-Disease OCT Image Classification Datasets with Explainable AI**, investigates the efficacy of deep learning and hybrid models for classifying retinal diseases using Optical Coherence Tomography (OCT) images. OCT is a pivotal non-invasive imaging technique that provides high-resolution retinal visualizations, crucial for diagnosing conditions like diabetic retinopathy, age-related macular degeneration, and glaucoma, which contribute significantly to global visual impairment. The research addresses challenges in manual OCT diagnosis, such as subjectivity, time consumption, and limited accessibility, by leveraging artificial intelligence (AI) to enhance diagnostic accuracy and scalability. Five models were evaluated—LBP-SVM, GoogleNet-RandomForest, ResNet50-Pretrained, Hybrid MobileNet-ViT (Pipelined), and Hybrid MobileNet-ViT (Attention Weighted)—across three datasets: Kermany OCT, OCTDL, and MMC-AMD. These datasets encompass diverse retinal pathologies, enabling robust assessment of model performance under varying conditions. The hybrid MobileNet-ViT models, combining convolutional neural networks (CNNs) for local feature extraction with Vision Transformers (ViTs) for global context modeling, achieved superior accuracies of up to 99.48% on the Kermany dataset, 93.53% on OCTDL, and 72.52% on MMC-AMD, outperforming traditional CNN-based models. Explainability techniques, including Grad-CAM and Integrated Gradients, were employed to enhance model transparency, highlighting critical image regions influencing classifications and fostering clinical trust. The study reveals that hybrid models offer improved generalization and robustness, particularly in handling dataset variability, though challenges like class imbalance and feature overlap persist. Future work should focus on refining attention mechanisms, optimizing feature extraction, and integrating contrastive learning to further enhance classification accuracy and interpretability, paving the way for reliable, scalable, and clinically viable AI-driven OCT diagnostic systems.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Artificial Intelligence in Ophthalmic Diagnostics . . . 2
  - 1.2.1 Challenges in Processing Retinal OCT Images Compared to General Images . . . 2
- 1.3 Problem Statement . . . 3

2 Related Work **4**
- 2.1 Retinal Diseases . . . 4
  - 2.1.1 Kermany OCT Dataset . . . 4
  - 2.1.2 OCTDL Dataset . . . 5
  - 2.1.3 MMC-AMD Dataset . . . 5
- 2.2 ML/DL Techniques . . . 6
  - 2.2.1 Local Binary Pattern (LBP) for Feature Extraction in OCT Imaging . . . 6
    - 2.2.1.1 LBP Computation . . . 6
  - 2.2.2 MobileNetV3 . . . 7
  - 2.2.3 Encoder and Decoder Architecture . . . 8
  - 2.2.4 Recurrent Neural Networks (RNN) . . . 8
  - 2.2.5 Long Short-Term Memory (LSTM) . . . 9
  - 2.2.6 Attention Mechanisms in Deep Learning . . . 9
    - 2.2.6.1 Scaled Dot-Product Attention and Multi-Head Attention . . . 10
    - 2.2.6.2 Attention in Transformers and Vision Transformers . . . 10
  - 2.2.7 Transformers . . . 11
  - 2.2.8 Vision Transformers (ViTs) . . . 12
  - 2.2.9 Grad-CAM: Gradient-weighted Class Activation Mapping . . . 13
  - 2.2.10 Integrated Gradients (IG) . . . 14
- 2.3 Studies on Deep Learning and Hybrid Models for OCT Classification 14
- 2.4 Summary . . . 16

3 Proposed Methodology **18**
- 3.1 Hybrid : MobileNetV3 + Vision Transformer (ViT) (Pipelined) . . . 18
  - 3.1.1 Feature Extraction & Transformation . . . 18
  - 3.1.2 Vision Transformer Processing & Patch Embedding . . . 18
  - 3.1.3 Self-Attention & Classification . . . 19
- 3.2 Hybrid MobileNet-ViT (Attention Weighted) . . . 20

4 Results and Discussion **22**
4.1 Dataset Kermany OCT . . . 22
4.1.1 Model A: LBP-SVM . . . 22
4.1.1.1 Confusion Matrix Analysis . . . 23
4.1.1.2 Discussion and Future Improvements . . . 23
4.1.2 Model B: GoogleNet-RandomForest . . . 24
4.1.2.1 Confusion Matrix Analysis . . . 24
4.1.2.2 Discussion and Future Work . . . 25
4.1.3 Model C: ResNet50-Pretrained . . . 25
4.1.3.1 Confusion Matrix Analysis . . . 25
4.1.3.2 Discussion and Future Work . . . 26
4.1.4 Model D: Hybrid MobileNet-ViT (Pipelined) . . . 26
4.1.4.1 Confusion Matrix Analysis . . . 27
4.1.4.2 Prediction Explanation Using Integrated Gradients . . . 27
4.1.4.3 Discussion and Future Work . . . 27
4.1.5 Model E: Hybrid MobileNet-ViT (Attention Weighted) . . . 28
4.1.5.1 Confusion Matrix Analysis . . . 28
4.1.5.2 Prediction Explanation Using Grad-CAM . . . 29
4.1.5.3 Discussion and Future Work . . . 29
4.2 OCTDL Dataset . . . 29
4.2.1 Model A: LBP-SVM . . . 30
4.2.1.1 Confusion Matrix Analysis . . . 30
4.2.1.2 Discussion and Future Work . . . 31
4.2.2 Model B: GoogleNet-RandomForest . . . 31
4.2.2.1 Confusion Matrix Analysis . . . 31
4.2.2.2 Discussion and Future Work . . . 32
4.2.3 Model C: ResNet50-Pretrained . . . 32
4.2.3.1 Confusion Matrix Analysis . . . 33
4.2.3.2 Discussion and Future Work . . . 34
4.2.4 Model D: Hybrid MobileNet-ViT (Pipelined) . . . 34
4.2.4.1 Confusion Matrix Analysis . . . 34
4.2.4.2 Prediction Explanation Using Integrated Gradients . . . 35
4.2.4.3 Discussion and Future Work . . . 36
4.2.5 Model E: Hybrid MobileNet-ViT (Attention Weighted) . . . 36
4.2.5.1 Confusion Matrix Analysis . . . 36
4.2.5.2 Prediction Explanation Using Grad-CAM . . . 37
4.2.5.3 Discussion and Future Work . . . 37
4.3 Dataset MMC-AMD . . . 38
4.3.1 Model A: LBP-SVM . . . 38
4.3.1.1 Confusion Matrix Analysis . . . 38
4.3.1.2 Discussion and Future Work . . . 39
4.3.2 Model B: GoogleNet-RandomForest . . . 39
4.3.2.1 Confusion Matrix Analysis . . . 40
4.3.2.2 Discussion and Future Work . . . 40
4.3.3 Model C: ResNet50-Pretrained . . . 41
4.3.3.1 Discussion and Future Work . . . 42
4.3.4 Model D: Hybrid MobileNet-ViT (Pipelined) . . . 42

4.3.4.1 Confusion Matrix Analysis 42
4.3.4.2 Prediction Explanation Using Integrated Gradients . 43
4.3.4.3 Discussion and Future Work 43
4.3.5 Model E: Hybrid MobileNet-ViT (Attention Weighted) 44
4.3.5.1 Confusion Matrix Analysis 44
4.3.5.2 Prediction Explanation Using Grad-CAM 45
4.3.5.3 Discussion and Future Work 45
4.4 Summary of Test Accuracies 45

Bibliography **50**

List of Figures

2.1 MobileNetV3 Architecture Overview: Bottleneck blocks, depthwise convolutions, SE layers, and hard-swish activations improve efficiency. 7
2.2 Attention Mechanism 10
2.3 The Transformer Model 11
2.4 Vision Transformer Model 12
2.5 Grad-CAM overview: The CNN extracts feature maps from the input image, while guided backpropagation refines feature importance. Grad-CAM generates a localization map, highlighting influential regions for classification. 13
2.6 Integrated Gradients for a FireBoat 15

3.1 Hybrid MobileNet-ViT Pipeline: MobileNetV3 extracts local features, transformed into global representations via ViT for final classification. 19
3.2 Hybrid MobileNet-ViT (Attention Weighted) Pipeline: MobileNetV3 extracts local features, ViT models global dependencies, and a fusion layer combines representations for final classification. 20

4.1 Confusion Matrix for LBP-SVM on Kermany Dataset 23
4.2 Confusion Matrix for GoogleNet-RandomForest on Kermany Dataset 24
4.3 Confusion Matrix for ResNet50-Pretrained on Dataset MMC-AMD . 26
4.4 Confusion Matrix for Hybrid MobileNet-ViT (Pipelined) on Kermany Dataset 27
4.5 Integrated Gradients for CNV disease - Kermany Dataset 27
4.6 Confusion Matrix for Hybrid MobileNet-ViT (Attention Weighted) on Kermany Dataset 28
4.7 GradCAM for CNV disease - Kermany Dataset 29
4.8 Confusion Matrix for LBP-SVM on OCTDL Dataset 30
4.9 Confusion Matrix for GoogleNet-RandomForest on OCTDL Dataset . 32
4.10 Confusion Matrix for ResNet50-Pretrained on OCTDL Dataset 33
4.11 Confusion Matrix for Hybrid MobileNet-ViT (Pipelined) on OCTDL Dataset 35
4.12 Integrated Gradients for RVO disease - OCTDL Dataset 35
4.13 Confusion Matrix for Hybrid MobileNet-ViT (Attention Weighted) on OCTDL Dataset 37
4.14 GradCAM for RVO disease - OCTDL Dataset 37
4.15 Confusion Matrix for LBP-SVM on MMC-AMD Dataset 39
4.16 Confusion Matrix for GoogleNet-RandomForest on MMC-AMD Dataset 40

4.17 Confusion Matrix for ResNet50-Pretrained on MMC-AMD Dataset . 41
4.18 Confusion Matrix for Hybrid MobileNet-ViT (Pipelined) on MMC-AMD Dataset 43
4.19 Integrated Gradients for PCV disease - MMC-AMD Dataset 43
4.20 Confusion Matrix for Hybrid MobileNet-ViT (Attention Weighted) on MMC-AMD Dataset 44
4.21 Grad-CAM visualization for Hybrid MobileNet-ViT (Attention Weighted): Activation regions influencing classification decisions. 45

List of Tables

4.1 Classification Report for LBP-SVM on Kermany Dataset 22
4.2 Classification Report for GoogleNet-RandomForest on Kermany Dataset 24
4.3 Classification Report for ResNet50-Pretrained on Kermany Dataset . 25
4.4 Classification Report for Hybrid MobileNet-ViT (Pipelined) on Kermany Dataset 26
4.5 Classification Report for Hybrid MobileNet-ViT (Attention Weighted) on Kermany Dataset 28
4.6 Classification Report for LBP-SVM on OCTDL Dataset 30
4.7 Classification Report for GoogleNet-RandomForest on OCTDL Dataset 31
4.8 Classification Report for ResNet50-Pretrained on OCTDL Dataset . . 33
4.9 Classification Report for Hybrid MobileNet-ViT (Pipelined) on OCTDL Dataset 34
4.10 Classification Report for Hybrid MobileNet-ViT (Attention Weighted) on OCTDL Dataset 36
4.11 Classification Report for LBP-SVM on MMC-AMD Dataset 38
4.12 Classification Report for GoogleNet-RandomForest on MMC-AMD Dataset 39
4.13 Classification Report for ResNet50-Pretrained on MMC-AMD Dataset 41
4.14 Classification Report for Hybrid MobileNet-ViT (Pipelined) on MMC-AMD Dataset 42
4.15 Classification Report for Hybrid MobileNet-ViT (Attention Weighted) on MMC-AMD Dataset 44
4.16 Final Test Accuracies of All Models Across Datasets 46

Chapter 1

Introduction

Optical Coherence Tomography (OCT) is a non-invasive imaging modality that provides high-resolution, cross-sectional images of the retina, enabling detailed visualization of its microstructures. It works based on low-coherence interferometry, which measures the echo time delay and intensity of backscattered light to generate tomographic images of biological tissues (1).

OCT has revolutionized ophthalmic diagnostics by allowing clinicians to detect and monitor retinal diseases with precision. Its resolution (typically 1–15 microns) is much higher than that of traditional imaging methods like fundus photography, and it facilitates in vivo retinal imaging in a matter of seconds. Over the past decade, OCT has become the gold standard for diagnosing and managing retinal and choroidal disorders, offering both quantitative and qualitative insights into retinal morphology. The widespread adoption of OCT has significantly improved early disease detection, aiding clinicians in personalized treatment planning.

1.1 Motivation

The global burden of visual impairment continues to rise, with eye diseases such as diabetic retinopathy (DR), age-related macular degeneration (AMD), and glaucoma contributing significantly to preventable blindness. According to the World Health Organization, over 2.2 billion people worldwide suffer from vision impairment, and at least 1 billion of these cases could have been prevented or remain unaddressed (2).

Early screening plays a critical role in halting disease progression, particularly for conditions like diabetic retinopathy, which often remain asymptomatic in initial stages. Studies indicate that systematic screening programs can reduce the risk of severe vision loss by up to 50% in diabetic populations (3). Despite these advancements, the manual diagnosis process still presents substantial challenges:

- Subjectivity: Diagnosis varies among practitioners due to differences in experience, training, and fatigue.
- Time Consumption: Manual inspection of retinal images is labor-intensive and not scalable for large screening programs.

- Limited Accessibility: Many rural and under-resourced areas lack trained ophthalmologists, leading to delayed diagnoses.

A study by Gulshan et al. (4) revealed that trained ophthalmologists exhibit significant inter-grader variability in diabetic retinopathy classification, further emphasizing the need for automated, scalable, and reliable diagnostic systems.

1.2 Artificial Intelligence in Ophthalmic Diagnostics

Artificial Intelligence (AI) has transformed medical imaging by improving diagnostic accuracy, reducing observer variability, and streamlining clinical workflows. AI-powered algorithms analyze large volumes of imaging data swiftly and consistently, providing valuable assistance to clinicians in making informed decisions.

In ophthalmology, AI models have demonstrated remarkable performance in automating disease detection and classification across fundus photography, fluorescein angiography, and OCT scans. Deep learning architectures, particularly Convolutional Neural Networks (CNNs), have proven effective at identifying retinal pathologies with accuracy comparable to human experts (5). However, traditional deep learning models still face limitations.

To address these concerns, researchers have explored hybrid models that integrate CNNs with Support Vector Machines (SVMs) or Transformers to improve generalization and interpretability (6). Additionally, explainability techniques such as Grad-CAM and Integrated Gradients are gaining traction in enhancing AI transparency (7).

1.2.1 Challenges in Processing Retinal OCT Images Compared to General Images

Processing retinal OCT images presents unique challenges compared to general images, such as natural photographs of objects or landscapes, due to their distinct structural and pathological characteristics. These challenges necessitate specialized AI approaches, as outlined below:

- **High-Resolution Layered Structure**: OCT images capture cross-sectional views of the retina with a resolution of 1–15 microns, revealing intricate layered structures, such as the inner nuclear layer, outer plexiform layer, and retinal pigment epithelium (RPE). Unlike general images, where objects are defined by broad color and texture variations, OCT images require models to discern subtle layer disruptions or thickness variations indicative of diseases like diabetic macular edema (DME) or choroidal neovascularization (CNV). This demands precise feature extraction, as small pixel-level changes can be diagnostically significant (1).

- **Subtle Pathological Features**: Retinal diseases often manifest as minute anomalies, such as hyper-reflective drusen deposits or micro-cystic spaces in DME, which are less conspicuous than the prominent features in general images (e.g., edges of a tree or a car's shape). Detecting these subtle features requires models with high sensitivity to fine-grained textures and patterns, increasing the risk of overfitting or missing critical diagnostic cues if not properly tuned (8).

- **Domain-Specific Noise and Artifacts**: OCT images are prone to noise, such as speckle noise from low-coherence interferometry, and artifacts from patient movement or imaging errors. These artifacts can obscure pathological features, unlike general images, where noise (e.g., lighting variations) is often less critical to object recognition. Robust preprocessing and noise-robust models are essential to ensure accurate OCT classification (9).

- **Need for Global and Local Context**: OCT image analysis demands both local feature analysis (e.g., fluid texture in CNV) and global structural understanding (e.g., retinal layer organization in DME). General images often prioritize local features (e.g., texture of a flower) without requiring extensive long-range dependencies. The hybrid MobileNet-ViT model proposed in this study addresses this by combining CNNs for local features with Transformers for global context, a necessity driven by OCT's unique requirements (10).

These challenges underscore the need for advanced AI architectures tailored to OCT's complexities. By integrating hybrid models and explainability techniques, this study aims to overcome these obstacles, enhancing the accuracy, robustness, and clinical applicability of automated OCT diagnostics.

1.3 Problem Statement

While deep learning has significantly improved OCT-based retinal disease classification, single-model architectures often struggle with generalization due to dataset variability, model complexity, and interpretability concerns. There is a critical need to evaluate how hybrid models compare to traditional deep learning approaches in terms of accuracy, robustness, and explainability.

This study, titled Comparative Analysis of Deep Learning and Hybrid Models on Multi-Disease OCT Image Classification Datasets with Explainable AI, aims to:

- Assess the comparative performance of CNNs, Vision Transformers, and hybrid models across diverse OCT datasets.
- Analyze the impact of dataset variability on model robustness and adaptability.
- Investigate the effectiveness of explainability techniques in enhancing model transparency for clinical applications.
- Explore strategies for improving generalization through transfer learning and domain adaptation.

By systematically examining these aspects, this research contributes to the development of more interpretable, reliable, and clinically viable AI-powered OCT diagnostic systems, bridging the gap between deep learning advancements and real-world medical deployment.

Chapter 2

Related Work

2.1 Retinal Diseases

This section provides an overview of the retinal diseases covered in the Kermany OCT, OCTDL, and MMC-AMD datasets, focusing on their distinguishing features in Optical Coherence Tomography (OCT) images. OCT, a non-invasive imaging modality, captures high-resolution cross-sectional views of the retina, enabling precise visualization of pathological changes critical for diagnosing retinal disorders. The diseases are organized by dataset, with their distinguishing features described to highlight the textural and structural patterns that differentiate them.

2.1.1 Kermany OCT Dataset

The Kermany OCT dataset includes four classes: Normal, Choroidal Neovascularization (CNV), Diabetic Macular Edema (DME), and Drusen, each exhibiting distinct OCT features.

- **Normal**: Normal retinas display well-organized layers with smooth, regular contours and no pathological disruptions. The inner nuclear layer, outer plexiform layer, and retinal pigment epithelium (RPE) are clearly delineated with uniform thickness and reflectivity, distinguishing them from diseased states with fluid or deposits (**?**).

- **Choroidal Neovascularization (CNV)**: CNV is characterized by subretinal or intraretinal fluid accumulation, hyper-reflective material beneath or within the RPE, and disruption of the RPE and Bruch's membrane. Irregular RPE elevation with fluid and hyper-reflective spots, often spanning multiple retinal regions, distinguishes CNV from localized anomalies like drusen (1).

- **Diabetic Macular Edema (DME)**: DME presents with intraretinal cysts or fluid-filled spaces, diffuse retinal thickening, and hyper-reflective hard exudates. Scattered cystic spaces altering the global retinal curvature differentiate DME from CNV's more focal disruptions (3).

- **Drusen**: Drusen appear as small, hyper-reflective deposits beneath the RPE, causing dome-shaped elevations with minimal fluid. Their localized, discrete nature with

preserved overlying retinal layers sets them apart from CNV's fluid-heavy patterns or DME's cystic edema (**?**).

2.1.2 OCTDL Dataset

The OCTDL dataset encompasses seven classes, likely including Normal, CNV, DME, Drusen, Retinal Vein Occlusion (RVO), Central Serous Retinopathy (CSR), and Vitreomacular Traction (VMT), based on common OCT pathologies and model performance analysis.

- **Normal**: As in the Kermany dataset, normal retinas show uniform layer organization without fluid, deposits, or traction, serving as a baseline for pathological comparisons.
- **Choroidal Neovascularization (CNV)**: CNV exhibits similar features to those in Kermany, with fluid and hyper-reflective material disrupting the RPE, distinguished by its irregular, multi-regional impact.
- **Diabetic Macular Edema (DME)**: DME's cystic spaces and retinal thickening are consistent across datasets, with its global curvature disruption setting it apart from localized conditions like drusen.
- **Drusen**: Drusen's hyper-reflective deposits and minimal layer disruption remain distinct, with their dome-shaped RPE elevations contrasting with fluid-heavy diseases.
- **Retinal Vein Occlusion (RVO)**: RVO is marked by intraretinal edema with cystoid spaces, retinal hemorrhages, and inner retinal thickening, often with serous detachment. Its prominent inner retinal swelling and unilateral distribution differentiate it from DME's diffuse cysts or CNV's subretinal fluid (11).
- **Central Serous Retinopathy (CSR)**: CSR features serous detachment of the neurosensory retina with subretinal fluid and thickened choroid, sparing intraretinal layers. Its smooth, dome-shaped detachment contrasts with RVO's inner retinal edema or CNV's hyper-reflective disruptions (1).
- **Vitreomacular Traction (VMT)**: VMT shows tractional distortion of the macula with foveal tenting, possible intraretinal cysts, and preserved outer layers. Its focal traction distinguishes it from diffuse edema in DME or deposits in drusen (1).

2.1.3 MMC-AMD Dataset

The MMC-AMD dataset includes four classes: Normal, Dry Age-Related Macular Degeneration (Dry AMD), Wet Age-Related Macular Degeneration (Wet AMD), and Polypoidal Choroidal Vasculopathy (PCV).

- **Normal**: Normal retinas exhibit consistent layer organization, free of atrophy, fluid, or neovascular changes, as in other datasets.

- **Dry Age-Related Macular Degeneration (Dry AMD)**: Dry AMD is characterized by drusen deposits, RPE atrophy, and outer retinal thinning, with geographic atrophy in advanced cases. Its atrophic changes without significant fluid distinguish it from wet AMD or CNV (12).
- **Wet Age-Related Macular Degeneration (Wet AMD)**: Wet AMD presents with subretinal or intraretinal fluid, choroidal neovascular membranes, hemorrhages, and RPE detachment. Its aggressive neovascular activity and fluid differentiate it from dry AMD's atrophic patterns (12).
- **Polypoidal Choroidal Vasculopathy (PCV)**: PCV features polypoidal lesions beneath the RPE, subretinal fluid or hemorrhage, and notched RPE detachment with thickened choroid. Its polyp-like structures and notched detachments set it apart from CNV's diffuse fluid or wet AMD's neovascular membranes (12).

These distinguishing features, ranging from localized deposits to global layer disruptions, underscore the complexity of OCT-based diagnosis. The Hybrid MobileNet-ViT model leverages these patterns, combining local feature extraction with global context modeling, to achieve robust classification across diverse datasets, as discussed in Chapter **Results and Discussions**.

2.2 ML/DL Techniques

2.2.1 Local Binary Pattern (LBP) for Feature Extraction in OCT Imaging

Local Binary Pattern (LBP) is a texture descriptor used in image processing to characterize patterns within local neighborhoods of pixels. Initially introduced for texture classification, LBP has gained prominence in medical imaging due to its ability to capture fine-grained spatial variations. It is particularly useful for analyzing Optical Coherence Tomography (OCT) images, where subtle retinal texture differences indicate various pathological conditions.

2.2.1.1 LBP Computation

The LBP operator works by comparing each pixel's intensity value with its surrounding neighbors, encoding structural information as a binary pattern.

Given a central pixel p with intensity I_p, its surrounding P neighborhood pixels are compared using:

$$LBP_p = \sum_{i=0}^{P-1} s(I_i - I_p) \cdot 2^i \tag{2.1}$$

where:

- I_p is the intensity of the central pixel.
- I_i represents the neighboring pixel intensities.

- The function $s(x)$ is defined as:

$$s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{2.2}$$

- The binary pattern is converted into a decimal value using powers of 2.

2.2.2 MobileNetV3

MobileNetV3 is a deep learning model optimized for mobile and edge devices, balancing accuracy and computational efficiency through efficient convolutional operations, lightweight attention mechanisms, and non-linear activations. It integrates several improvements over its predecessors, including depthwise separable convolutions, Squeeze-and-Excite (SE) layers for channel-wise recalibration, Hard-Swish activation for enhanced non-linearity, and lightweight bottleneck blocks with residual connections. Neural Architecture Search (NAS) further optimizes layer configurations for mobile platforms.

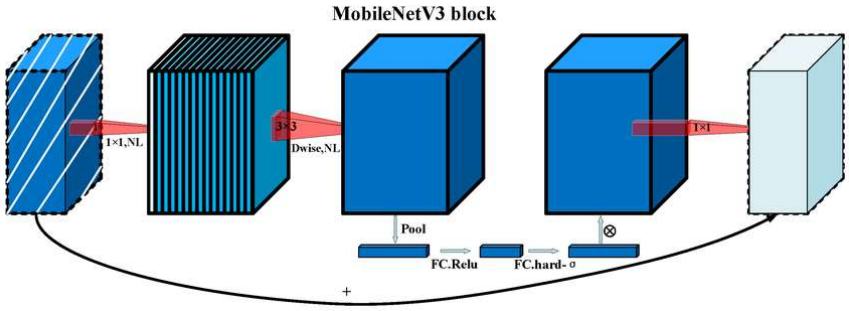

Figure 2.1: MobileNetV3 Architecture Overview: Bottleneck blocks, depthwise convolutions, SE layers, and hard-swish activations improve efficiency.

MobileNetV3 reduces computational cost by replacing standard convolutions with depthwise separable convolutions, where the total operations are:

$$\text{Computational Cost} = k^2 \cdot C_{\text{in}} + C_{\text{in}} \cdot C_{\text{out}} \tag{2.3}$$

where k is the kernel size and $C_{\text{in}}, C_{\text{out}}$ are input and output channels. SE layers refine channel-wise attention using an excitation function:

$$s = \sigma(W_2 \cdot \text{ReLU}(W_1 \cdot g)) \tag{2.4}$$

where g represents global average pooled features, W_1, W_2 are learnable weight matrices, and σ is the sigmoid activation.

To further improve efficiency, MobileNetV3 adopts Hard-Swish activation:

$$h\text{swish}(x) = x \cdot \frac{\text{ReLU6}(x+3)}{6} \tag{2.5}$$

which enhances non-linearity while maintaining smooth gradient flow.

Overall, MobileNetV3 achieves superior performance compared to its predecessor, MobileNetV2, offering optimized feature extraction, lower computational cost, and improved efficiency for edge AI applications.

2.2.3 Encoder and Decoder Architecture

The Encoder-Decoder model is an important deep learning framework widely used in natural language processing, machine translation, and speech recognition. It employs two components, an encoder and a decoder, to achieve sequence-to-sequence conversion by encoding the input sequence and generating the decoding of the target sequence.

The model consists of two main components:

- **Encoder**: Converts the input sequence into a fixed-dimensional vector representation, capturing the semantic information of the sequence. It iterates over each element in the input sequence, updating its hidden state to obtain a comprehensive representation.
- **Decoder**: Gradually generates an output sequence based on the encoded vector representation. It uses the encoder's output as the initial hidden state and generates each subsequent output based on the previous time step's output and the current hidden state.

The output of a decoder can be:

- A sequence of discrete symbols (e.g., words in machine translation).
- A continuous value (e.g., speech signals in voice processing).

This architecture enables handling variable-length input and output sequences, incorporating contextual understanding and generation capabilities at the sequence level.

2.2.4 Recurrent Neural Networks (RNN)

Recurrent Neural Networks (RNNs) are specialized neural network models designed for sequential data processing, allowing information from previous time steps to influence the current step. This temporal dependency makes RNNs particularly effective for tasks such as natural language processing, speech recognition, and machine translation.

A standard RNN unit consists of an input layer, a hidden layer with reflexive connections enabling information transfer across time steps, and an output layer. By linking multiple RNN units, a complete RNN model can be built, enabling sequential pattern recognition.

One of the major challenges in traditional RNNs is the issue of vanishing and exploding gradients during training, which limits their ability to learn long-term dependencies. The **Real-Time Recurrent Learning** (RTRL) algorithm was introduced to address these challenges by enabling precise gradient computation for continuous data streams. In modern applications, RNN-based encoder-decoder architectures have gained popularity, where the encoder converts an input sequence into a fixed-length representation and the decoder generates an output sequence.

RNNs process sequence data recursively, using the output of the previous time step as input for the next.

RNNs offer efficient handling of sequential data, modeling long-term dependencies, and supporting variable-length inputs and outputs. However, they suffer from slow training speeds and difficulty in learning extended dependencies due to gradient-related issues. To mitigate these limitations, advanced models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been developed, offering improved performance in sequence-based tasks.

2.2.5 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a variant of Recurrent Neural Networks (RNNs) designed to address the gradient vanishing and explosion problems common in standard RNNs. Traditional RNNs struggle to retain long-term dependencies due to loss of earlier sequence information over extended time steps. This limitation affects their ability to model long-sequence patterns, making them unsuitable for tasks requiring sustained memory.

Hochreiter and Schmidhuber introduced LSTMs, which incorporate specialized mechanisms known as gates to regulate information flow. These gates—forget, input, and output—allow the model to selectively retain or discard information, thereby mitigating the challenges of standard RNNs. The forget gate determines which past information should be discarded based on a sigmoid function, ensuring efficient memory control. The input gate identifies relevant new information and updates the cell state, using a combination of sigmoid and tanh activations to refine stored representations. Finally, the output gate determines which parts of the cell state contribute to the next step, regulating the flow of information in sequence generation.

LSTMs offer several advantages over conventional RNNs, including enhanced capability in modeling long-term dependencies, effective mitigation of vanishing gradients, and support for variable-length input sequences. These properties make them widely applicable in fields such as natural language processing, speech recognition, stock market prediction, and biomedical time-series analysis. Their ability to preserve context over extended durations makes them particularly valuable for sequential tasks.

Despite their advantages, LSTMs suffer from slow training speeds and high computational costs due to their complex gating mechanisms. As an alternative, Gated Recurrent Units (GRUs) provide a simplified architecture that achieves comparable accuracy while reducing computational overhead. Nevertheless, LSTMs remain foundational in deep learning models where long-term sequence retention is crucial.

2.2.6 Attention Mechanisms in Deep Learning

Attention mechanisms have become a fundamental component of modern deep learning models, enabling them to focus selectively on relevant parts of input sequences. An attention function maps a query and a set of key-value pairs to an output, where all components are represented as vectors. The output is computed as a weighted sum of the values, and the weight assigned to each value is derived from a compatibility function that compares the query with the corresponding key.

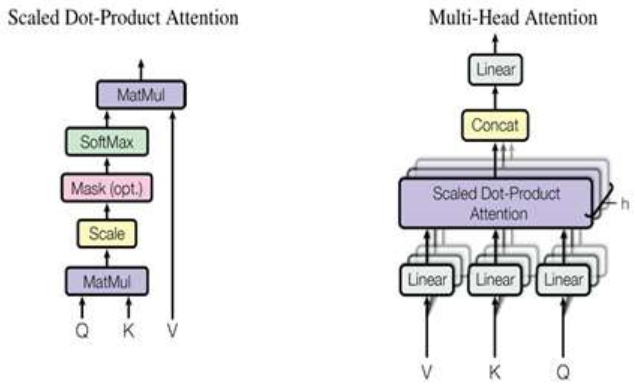

Figure 2.2: Attention Mechanism

2.2.6.1 Scaled Dot-Product Attention and Multi-Head Attention

Attention mechanisms enhance a model's ability to weigh different parts of the input sequence based on relevance. The Scaled Dot-Product Attention computes attention scores using the dot product between query (Q) and key (K) vectors, normalized by the scaling factor d_k:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.6}$$

where Q represents the current token's representation, K contains all tokens in the sequence, and V holds the relevant information. Scaling ensures numerical stability by preventing extreme softmax outputs that could lead to vanishing gradients.

Multi-Head Attention improves representation learning by processing multiple attention heads in parallel, each capturing different relationships in the sequence:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.7}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W^O \tag{2.8}$$

where W_i^Q, W_i^K, W_i^V are learned projection matrices for each attention head, and W^O is an output projection matrix. Multi-Head Attention enables the model to attend to different parts of the input simultaneously, improving expressiveness, computational efficiency, and reducing information loss compared to single-head attention.

By integrating Scaled Dot-Product and Multi-Head Attention mechanisms, deep learning models efficiently capture dependencies in sequential data, benefiting tasks such as natural language processing and computer vision.

2.2.6.2 Attention in Transformers and Vision Transformers

Originally designed for natural language processing (NLP) tasks, Transformers revolutionized sequence modeling by discarding recurrent structures in favor of attention mech-

anisms. Their success led to adaptations for computer vision, giving rise to Vision Transformers (ViTs).

ViTs modify the standard Transformer architecture to process image data effectively. Instead of sequential token embeddings, images are divided into patch embeddings, which are then processed similarly to text tokens using attention mechanisms.

2.2.7 Transformers

Transformers have revolutionized deep learning architectures, significantly improving performance in natural language processing (NLP), image analysis, and speech recognition.

Traditional RNNs suffer from vanishing gradients in long sequences, reducing their effectiveness in capturing distant dependencies. Transformers overcome these limitations using self-attention mechanisms, allowing the model to selectively focus on relevant parts of the input while enabling parallel computation.

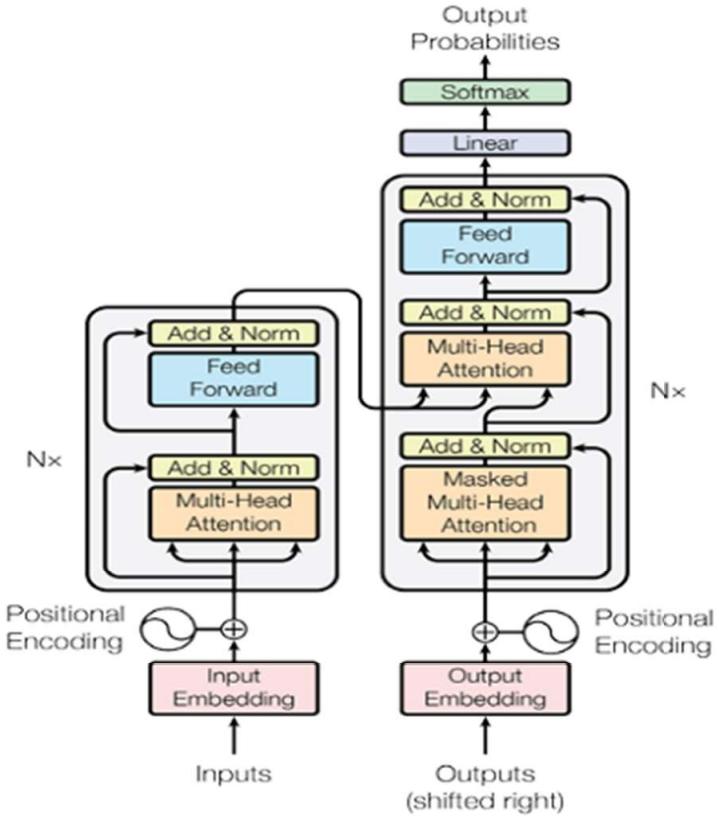

Figure 2.3: The Transformer Model

The core Transformer model consists of an encoder-decoder structure for sequence transduction tasks such as machine translation. The encoder comprises multiple identical layers incorporating multi-head self-attention to learn dependencies across sequence positions and position-wise feed-forward networks for feature transformation, with residual connections and layer normalization ensuring stability. The decoder follows the same layered structure but includes an additional multi-head attention mechanism over the encoder's output. To maintain causal dependencies, self-attention in the decoder is masked,

ensuring predictions at each position depend only on prior outputs.

Transformers outperform RNNs by leveraging parallel computation, effectively capturing long-range dependencies without suffering from vanishing gradients, and offering improved scalability. Their success has extended beyond NLP to areas like image classification, leading to the development of Vision Transformers (ViTs).

2.2.8 Vision Transformers (ViTs)

Vision Transformers (ViTs) adapt the standard Transformer architecture for image data, offering a competitive alternative to Convolutional Neural Networks (CNNs). Unlike **CNNs, which rely on spatial hierarchies**, ViTs process images using self-attention mechanisms, allowing them to capture **long-range dependencies** across an entire image.

ViTs first tokenize images into fixed-sized patches and embed them into a vector space. The attention mechanism compares token embeddings to compute alignment scores, determining their semantic and contextual relevance. Initially, each token's position is based solely on general meaning, but successive encoder layers refine this positioning to reflect deeper relationships.

Since Transformers do not inherently capture positional order, positional embeddings are introduced during pretraining to maintain spatial coherence. These embeddings converge into structured vector spaces where neighboring positions exhibit similarity, preserving spatial integrity for both NLP and image classification tasks.

A special classification token (CLS token) is appended at the beginning of the input sequence, aggregating information from all patches via self-attention. At the final Transformer layer, this token is extracted and passed to a classifier, enabling accurate image recognition.

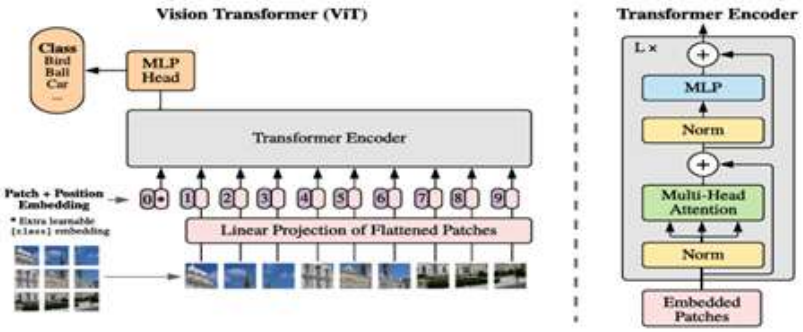

Figure 2.4: Vision Transformer Model

ViTs improve computational efficiency by eliminating inductive biases associated with CNNs, making them well-suited for large-scale visual tasks. Their ability to handle spatial dependencies effectively has positioned them as a powerful approach in computer vision (10).

2.2.9 Grad-CAM: Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) is a visualization technique used to interpret convolutional neural networks (CNNs) by identifying important image regions that influence the model's decision. It highlights relevant areas based on gradient information flowing into the last convolutional layer, generating heatmaps that enhance explainability.

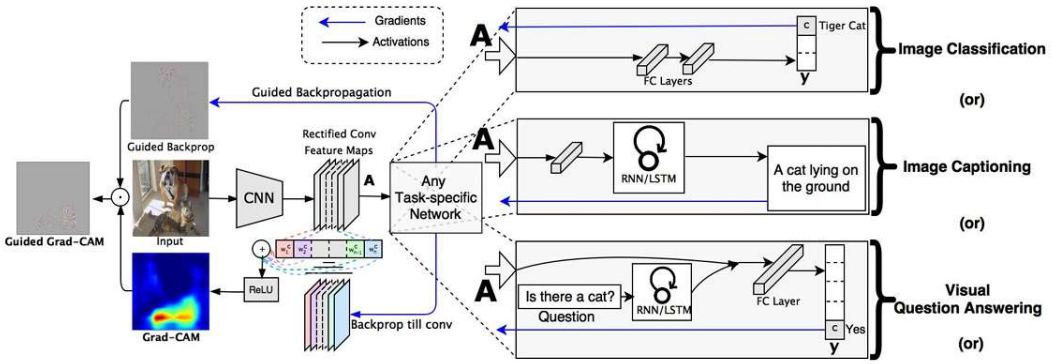

Figure 2.5: Grad-CAM overview: The CNN extracts feature maps from the input image, while guided backpropagation refines feature importance. Grad-CAM generates a localization map, highlighting influential regions for classification.

Grad-CAM computes a heatmap by weighting feature maps based on their gradients with respect to a target class. For a given CNN model, let A^k denote the activations of the last convolutional layer, and Y^c be the predicted score for class c. The gradient $\frac{\partial Y^c}{\partial A^k}$ measures the impact of each activation on the class prediction. Importance weights α_k^c are obtained using global average pooling:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{2.9}$$

where Z is the total number of pixels in the feature map. The final class activation map is computed as:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{2.10}$$

followed by normalization and upsampling to match the input image dimensions:

$$H^c = \text{Normalize}(L_{\text{Grad-CAM}}^c) \tag{2.11}$$

The resulting heatmap is superimposed onto the original image to highlight influential regions. Grad-CAM enhances model transparency across various applications, including image classification, visual question answering (VQA), and image captioning. In medical imaging, it plays a critical role in OCT-based disease classification, validating deep learning models for ophthalmology. Combined with Integrated Gradients, Grad-CAM

further strengthens interpretability, ensuring explainability in clinical AI deployments. By visualizing significant features influencing predictions, Grad-CAM fosters trust and transparency in AI-driven diagnostics (7).

2.2.10 Integrated Gradients (IG)

Integrated Gradients (IG) is an attribution method that assigns importance scores to input features based on model predictions, offering interpretability by quantifying each feature's contribution. Gradients determine how much a model's weights should be adjusted during backpropagation, providing insight into feature relevance. Standard gradient-based attribution methods suffer from noise artifacts, particularly when gradients are close to zero due to a flat function surface around well-trained inputs:

$$\text{gradient} = \frac{\partial \text{error}}{\partial \text{input}} = 0 \tag{2.12}$$

IG mitigates this issue by integrating gradients accumulated over multiple interpolated inputs, ensuring sensitivity and reducing noise. Formally, given a function $F : \mathbb{R}^n \to [0, 1]$, let x be the input and x' be a baseline (e.g., a black image for CNNs or a zero embedding for text models). IG computes the path integral of gradients along the straightline interpolation from x' to x:

$$\text{IntegratedGrad}_i(x) = (x_i - x'_i) \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{2.13}$$

where $\frac{\partial F(x)}{\partial x_i}$ represents the gradient of $F(x)$ with respect to the i-th dimension. The integral aggregates gradients over scaled inputs, improving robustness.

Two essential axioms ensure reliability in attribution methods:

1. **Sensitivity**: If an input and baseline differ in one feature, and their predictions change, the differing feature should have a non-zero attribution.
2. **Implementation Invariance**: Functionally equivalent models should produce consistent attribution results despite differing implementations.

IG combines the implementation invariance of gradients with the sensitivity of layer-wise relevance propagation (LRP) and DeepLift. It is widely applied in medical imaging, natural language processing, and finance, ensuring the clinical and ethical viability of AI models (13).

2.3 Studies on Deep Learning and Hybrid Models for OCT Classification

Several studies have explored the use of convolutional neural networks (CNNs) for classifying retinal diseases using OCT images. A landmark study by Kermany et al. (8) demonstrated that a CNN could classify images into CNV, DME, Drusen, and Normal categories with high accuracy, setting a foundational benchmark for future research. This

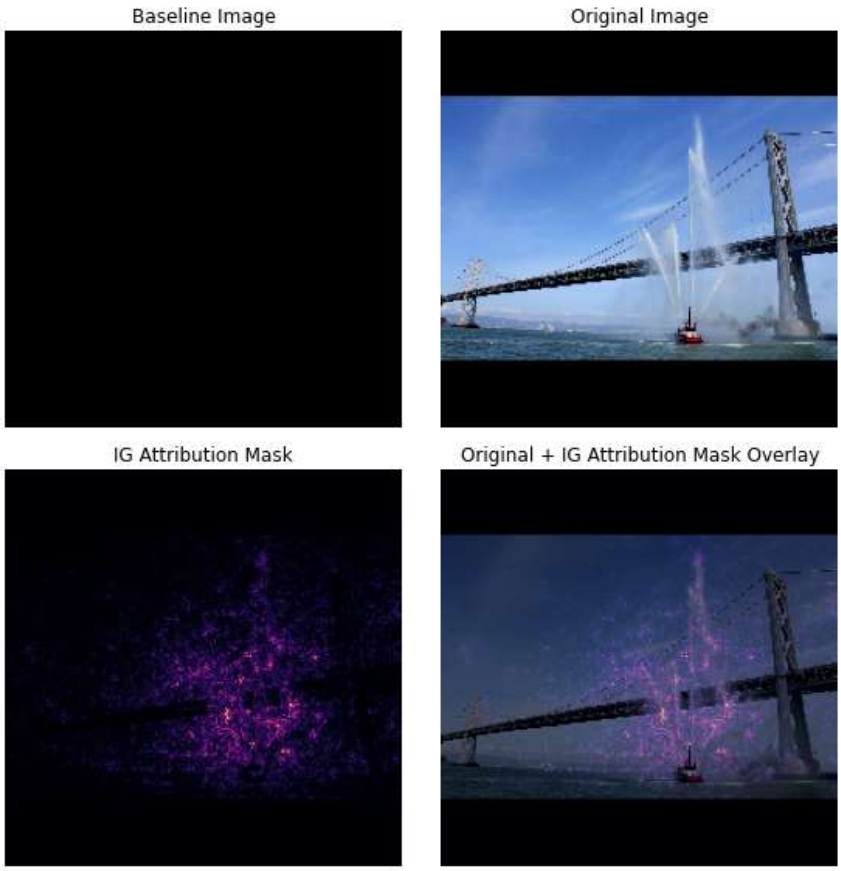

Figure 2.6: Integrated Gradients for a FireBoat

study used a transfer learning approach with InceptionV3, achieving 96.6% accuracy on the Kermany OCT dataset.

In another study, Kadry et al. (14) focused on age-related macular degeneration (AMD) classification using a combination of handcrafted features and deep learning models. They employed ResNet18 and achieved remarkable accuracy rates of 99.5% for dry AMD and 98.8% for wet AMD. This approach demonstrated the potential of integrating traditional feature extraction methods with deep learning for improved diagnostic accuracy.

Tasnim et al. (15) compared various deep learning models, including vanilla CNN, Xception, ResNet50, and MobileNetV2, for retinal disease detection using OCT images. Their results showed that MobileNetV2 achieved the highest accuracy of 99.17%, closely followed by the Xception model with 99.07%. These findings suggest that lightweight models like MobileNetV2 can offer high accuracy while being computationally efficient.

Recent studies have also explored Vision Transformers (ViT) for OCT image classification. Although initially designed for natural images, ViTs demonstrated competitive performance in medical imaging tasks when adapted through transfer learning and proper data augmentation, as shown by Dosovitskiy et al. (10) and further adapted for OCT in more recent works. Zhou et al. (16) demonstrated that ViTs outperformed CNNs in OCT classification tasks, particularly in cases where large-scale datasets were available.

However, despite high accuracy, these models often lack transparency and require

large amounts of labeled data, which has pushed researchers toward hybrid models and interpretability frameworks.

To overcome the limitations of purely CNN-based models, researchers have developed hybrid architectures that combine the strengths of multiple techniques. These models typically involve:

- Using CNNs for feature extraction and combining them with classifiers like Support Vector Machines (SVM), Random Forests, or Gradient Boosting to improve classification.
- Combining CNNs with Transformers, enabling both local feature learning (via convolution) and global context modeling (via attention mechanisms).

Jiang et al. (17) compared ViT-based models with state-of-the-art CNNs such as VGG16, ResNet50, DenseNet121, and EfficientNet for multi-class classification of AMD, diabetic macular edema (DME), and healthy eyes. The ViT models outperformed all CNN models, achieving an accuracy of 99.6%, demonstrating the efficacy of transformer-based architectures in medical imaging tasks.

Shen et al. (18) proposed a hybrid model combining CNNs and ViTs, which achieved an accuracy of 99% on the testing dataset. This performance surpassed that of standalone CNNs (accuracy ranging from 92% to 94%) and standalone transformers (accuracy ranging from 95% to 99%), indicating the advantage of hybrid models in capturing diverse feature representations.

Ma et al. (19) developed a hybrid ConvNet-Transformer network to classify OCT images into four categories: choroidal neovascularization (CNV), DME, drusen, and normal. Their hybrid model achieved accuracies ranging from 86% to 92%, outperforming standalone CNNs (76% to 90%) and standalone ViT models (79% to 86%) on two public retinal OCT datasets. Hybrid models also enable modularity, allowing components to be independently updated or interpreted, which is valuable in dynamic medical imaging tasks.

2.4 Summary

The theoretical concepts discussed in this chapter form the foundation for the proposed Hybrid MobileNet-ViT methodology, designed to address the challenges of Optical Coherence Tomography (OCT) image classification for retinal disease detection. This subsection elucidates how Local Binary Patterns (LBP), MobileNetV3, Vision Transformers (ViTs), attention mechanisms, and explainability techniques such as Grad-CAM and Integrated Gradients are integrated to create a robust, interpretable, and efficient model tailored for multi-disease OCT datasets.

Local Binary Patterns, as described in Section 2.1, provide a computationally efficient method for capturing fine-grained textural features in OCT images. While effective for traditional feature extraction, LBP alone struggles with the complex, hierarchical patterns inherent in retinal pathologies. To overcome this limitation, the proposed methodology leverages MobileNetV3 (Section 2.3) for its lightweight convolutional architecture, which excels at extracting local spatial features through depthwise separable convolutions and Squeeze-and-Excite layers. MobileNetV3's efficiency ensures that the model remains

computationally viable for deployment in resource-constrained clinical settings, a critical requirement for scalable OCT diagnostics.

However, convolutional neural networks (CNNs) like MobileNetV3 are limited in capturing long-range dependencies across an image, which are essential for modeling global structural relationships in retinal layers. Long-range dependencies refer to relationships between elements in an image (e.g., pixels, patches, or regions) that are spatially distant from each other. In image processing, these dependencies capture how features in one part of an image influence or correlate with features in another, far-away part. This is where Vision Transformers (ViTs), discussed in Section 2.9, play a pivotal role. ViTs process images as sequences of patches, using self-attention mechanisms (Section 2.7) to model global contextual dependencies. Global structural relationships refer to the overall organization and connectivity of features across an entire image, capturing the holistic structure rather than isolated parts. These relationships are essential for understanding the complete context of an image.

By integrating MobileNetV3 with ViT, the proposed hybrid model combines the strengths of local feature extraction (via CNNs) with global context modeling (via Transformers). The MobileNetV3 component extracts detailed spatial features, which are then transformed into patch embeddings for ViT to process, enabling the model to capture both localized textural anomalies and broader structural patterns indicative of retinal diseases.

The attention mechanisms detailed in Section 2.7, particularly Scaled Dot-Product and Multi-Head Attention, are central to the hybrid model's ability to weigh relevant image regions dynamically. In the Attention-Weighted variant of the Hybrid MobileNet-ViT, a fusion layer combines CNN-extracted features with ViT's global representations, using attention scores to prioritize disease-relevant regions. This approach enhances the model's ability to focus on pathological features, such as choroidal neovascularization or drusen, while suppressing irrelevant background noise, thereby improving classification accuracy.

Explainability is a cornerstone of the proposed methodology, addressing the interpretability concerns raised in Section 1.2. Grad-CAM (Section 2.10) and Integrated Gradients (Section 2.11) are employed to visualize the model's decision-making process. Grad-CAM generates heatmaps highlighting regions of the OCT image that influence classification, providing clinicians with intuitive insights into the model's focus areas. Integrated Gradients, by computing feature importance through gradient integration, offers a complementary perspective, quantifying the contribution of individual pixels to the prediction. These techniques ensure that the Hybrid MobileNet-ViT model is not only accurate but also transparent, fostering trust in clinical applications.

The integration of these concepts addresses key challenges in OCT image classification, such as dataset variability, computational complexity, and interpretability. By combining MobileNetV3's efficiency with ViT's global modeling capabilities, the hybrid model achieves robust generalization across diverse datasets like Kermany OCT, OCTDL, and MMC-AMD. The use of attention mechanisms enhances feature discriminability, while explainability techniques bridge the gap between AI predictions and clinical decision-making. This cohesive framework sets the stage for the proposed methodology detailed in Chapter 4, where the pipelined and attention-weighted variants of the Hybrid MobileNet-ViT model are further elaborated, demonstrating their superior performance in multi-disease OCT classification tasks.

Chapter 3

Proposed Methodology

3.1 Hybrid : MobileNetV3 + Vision Transformer (ViT) (Pipelined)

This architecture integrates MobileNetV3-Large for localized feature extraction and Vision Transformer (ViT-Base-16) for global context modeling, enhancing OCT disease classification.

3.1.1 Feature Extraction & Transformation

The input image tensor X is processed as:

$$X \in \mathbb{R}^{B\times 3\times 224\times 224}$$

where B represents the batch size. MobileNetV3 applies depth-wise separable convolutions, outputting feature maps:

$$F = MobileNet(X), \quad F \in \mathbb{R}^{B\times 960\times 7\times 7}$$

To reduce computational complexity, a 1×1 convolution compresses feature channels:

$$F' = Conv_{1\times 1}(F), \quad F' \in \mathbb{R}^{B\times 32\times 7\times 7}$$

Applying ReLU activation, followed by Adaptive Average Pooling, resizes the tensor:

$$F'' = Pool(ReLU(F')), \quad F'' \in \mathbb{R}^{B\times 32\times 224\times 224}$$

3.1.2 Vision Transformer Processing & Patch Embedding

ViT processes input as patch tokens, requiring channel alignment via a 16×16 convolution:

$$P = Conv_{patch}(F''), \quad P \in \mathbb{R}^{B\times 768\times 14\times 14}$$

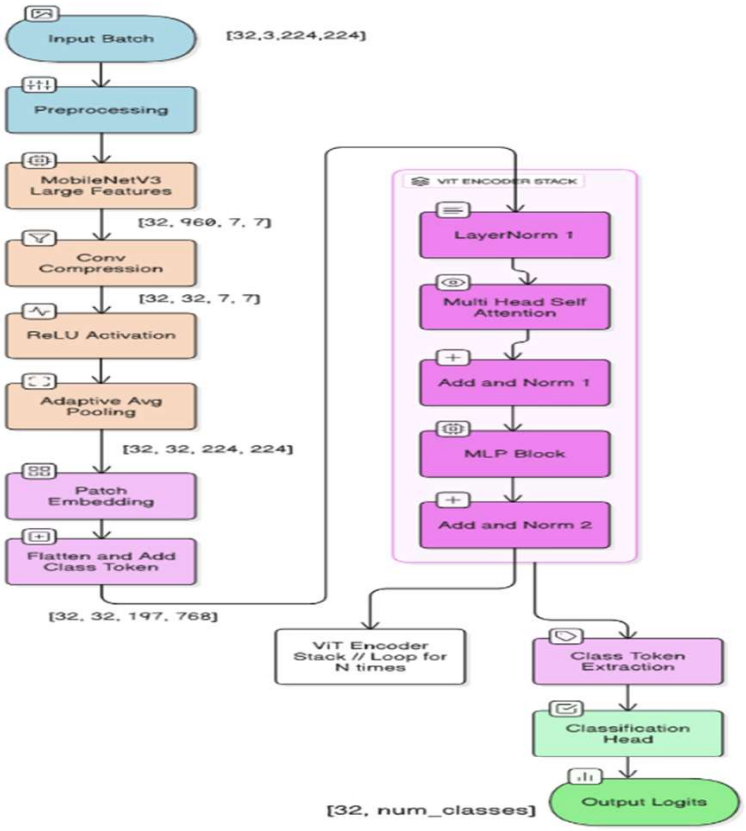

Figure 3.1: Hybrid MobileNet-ViT Pipeline: MobileNetV3 extracts local features, transformed into global representations via ViT for final classification.

where:

$$Conv_{patch} = Conv_{16\times16}(F'', 32 \to 768, stride = 16)$$

Each patch is flattened and concatenated with a learnable class token:

$$T = [P_{flatten}, CLS_{token}], \quad T \in \mathbb{R}^{B\times197\times768}$$

where 197 = 196 patches + 1 class token.

3.1.3 Self-Attention & Classification

The Transformer encoder stack applies multi-head self-attention layers:

$$T' = Transformer(T), \quad T' \in \mathbb{R}^{B\times197\times768}$$

Finally, the class token undergoes MLP-based classification:

$$C = Softmax(MLP(T_{cls})), \quad C \in \mathbb{R}^{B\times num_classes}$$

where Softmax activation outputs probabilities across 4 OCT disease categories.

By merging CNN-based local feature extraction with transformer-based global attention mechanisms, this hybrid model enhances interpretability and robustness in OCT-based disease classification.

3.2 Hybrid MobileNet-ViT (Attention Weighted)

The proposed hybrid architecture leverages the strengths of Convolutional Neural Networks (CNNs) for localized feature extraction and Vision Transformers (ViTs) for global contextual modeling, achieving robust classification of Optical Coherence Tomography (OCT) images. Specifically, it combines MobileNetV3-Large as a lightweight CNN backbone and ViT-Base-16 for capturing long-range dependencies.

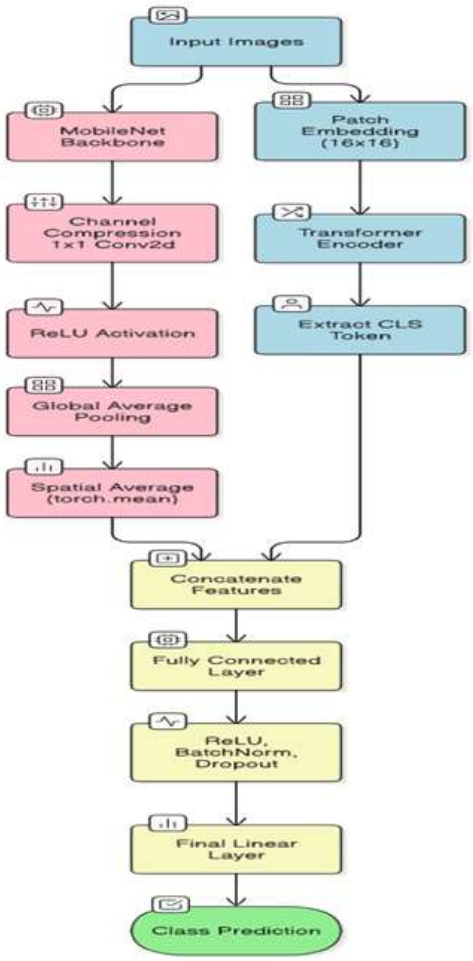

Figure 3.2: Hybrid MobileNet-ViT (Attention Weighted) Pipeline: MobileNetV3 extracts local features, ViT models global dependencies, and a fusion layer combines representations for final classification.

MobileNetV3 processes 224×224 RGB images, retaining only its mobilenet features module for extracting convolutional feature maps of shape $[B, 960, 7, 7]$. A 1×1 convolu-

tion reduces dimensions from 960 to 512, followed by global average pooling to produce a final feature vector $[B, 512]$:

$$F_{\text{CNN}} = \text{GlobalAvgPool}(\text{Conv2D}(F_{\text{prev}}, 512, 1, 1)) \tag{3.1}$$

Simultaneously, the ViT-Base-16 model divides the image into non-overlapping patches of size 16×16, yielding 196 patches per image. Each patch is embedded using a learnable class token, forming a sequence $[B, 197, 768]$, where only the class token $[B, 1, 768]$ represents global context:

$$F_{\text{ViT}} = T_{\text{CLS}} = \text{Conv2D}(F_{\text{input}}, 768, 16, 16) \tag{3.2}$$

The fusion process concatenates CNN-based localized features and ViT-based global embeddings along the feature dimension:

$$F_{\text{fusion}} = \text{Concat}(F_{\text{CNN}}, F_{\text{ViT}}) \quad (\text{Final Shape: } [B, 1280]) \tag{3.3}$$

This fused representation is projected into a 1024-dimensional space using a fully connected layer:

$$F_{\text{proj}} = \text{ReLU}(\text{BatchNorm}(\text{Linear}(F_{\text{fusion}}, 1024))) \tag{3.4}$$

A dropout layer (rate = 0.5) prevents overfitting before mapping the features to four disease categories via the final classification layer:

$$\text{Logits} = \text{Linear}(F_{\text{proj}}, 4) \tag{3.5}$$

Training utilizes cross-entropy loss, with backpropagation propagating gradients through ViT layers, modified patch embeddings, and the MobileNetV3 feature extractor. Pretrained initialization ensures faster convergence and strong generalization to OCT-specific retinal patterns.

Several key hyperparameters govern model behavior:

- Image Size: 224×224 (Balances computational efficiency and feature richness).
- Batch Size: 32 (Ensures stable gradient updates).
- Patch Size in ViT: 16×16 (Defines token granularity).
- Embedding Dimension: 768 (Captures global dependencies).

The Hybrid MobileNet-ViT (Attention Weighted) model aligns with emerging medical imaging research, combining CNNs for texture analysis and ViTs for structural dependencies, ensuring high accuracy and interpretability in OCT-based disease detection (16; 18).

Chapter 4

Results and Discussion

This chapter presents the results of the classification models applied to Optical Coherence Tomography (OCT) datasets, focusing on **classification reports**, **confusion matrices**, and overall **test accuracies**. The performance evaluation includes precision, recall, and F1-score for each disease class across different architectures. A summary of the final test accuracies for all models on all datasets is included in Table ??.

4.1 Dataset Kermany OCT

This section presents the classification report, confusion matrix, and performance analysis of multiple models on the **Kermany Dataset**, highlighting strengths and weaknesses observed during testing.

4.1.1 Model A: LBP-SVM

The classification report provides a detailed breakdown of precision, recall, and F1-score, evaluating how effectively the model distinguishes between different retinal disease categories.

Table 4.1: Classification Report for LBP-SVM on Kermany Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.500	0.521	0.510	242
Class 1	0.319	0.219	0.260	242
Class 2	0.364	0.413	0.387	242
Class 3	0.575	0.653	0.611	242
Overall Accuracy		45.14%		

The model achieves a **test accuracy of 45.14%**, highlighting classification challenges across disease categories. **Class 0 exhibits moderate performance**, whereas **Class 1 suffers from low recall (21.9%)**, leading to frequent misclassification. **Class 3 achieves the highest recall (65.3%) but struggles with precision, increasing false positives.** The performance gap between **Class 1 and Class 3** suggests **feature extraction limitations**, requiring refinement strategies.

4.1.1.1 Confusion Matrix Analysis

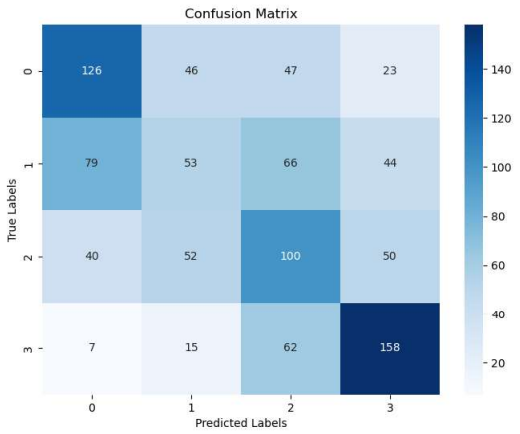

Figure 4.1: Confusion Matrix for LBP-SVM on Kermany Dataset

The confusion matrix further illustrates misclassification patterns:

- Class 0 shows moderate separability but some degree of misclassification.
- Class 1 suffers from significant overlap with other categories, negatively impacting recall.
- Class 2 and Class 3 demonstrate noticeable confusion, leading to frequent false positives.
- Feature optimization, such as dimensionality reduction or deep texture embeddings, could enhance classification accuracy.

4.1.1.2 Discussion and Future Improvements

The model's moderate accuracy emphasizes the need for feature refinement. Addressing the limitations in **Class 1 and Class 3** could **reduce recall issues** and enhance classification reliability. Improvements may include:

- **Strengthening LBP feature representations** through deeper texture embeddings.
- **Exploring hybrid classification frameworks**, integrating traditional SVM with CNN-based feature extraction.
- **Refining feature engineering techniques**, including multiscale pattern analysis and improved separability.

Future work should focus on optimizing feature extraction, incorporating CNN-based enhancements, and improving classification performance across OCT disease categories.

4.1.2 Model B: GoogleNet-RandomForest

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's performance in distinguishing retinal disease categories.

Table 4.2: Classification Report for GoogleNet-RandomForest on Kermany Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.941	0.864	0.901	242
Class 1	0.912	0.860	0.885	242
Class 2	0.941	0.996	0.968	242
Class 3	0.916	0.992	0.952	242
Overall Accuracy		92.77%		

The model achieves a **test accuracy of 92.77%**, demonstrating **strong classification capability**. **Class 0 exhibits high precision and recall**, confirming robust feature separability, while **Class 1 and Class 2 show balanced precision-recall trade-offs**, suggesting good generalization. **Class 3 attains the highest recall (99.2%) and strong precision (91.6%), reducing misclassification errors**

4.1.2.1 Confusion Matrix Analysis

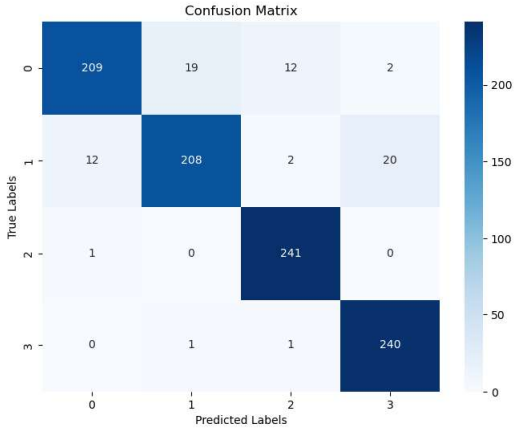

Figure 4.2: Confusion Matrix for GoogleNet-RandomForest on Kermany Dataset

The confusion matrix further illustrates classification performance:

- **Class 0 demonstrates minimal misclassification, confirming its separability.**
- **Class 1 and Class 2 show balanced performance, with minor prediction overlaps.**
- **Class 3 exhibits high recall, reinforcing its robustness in identifying pathological variations.**

- **Optimizing feature selection and ensemble learning could enhance precision and reduce residual prediction errors.**

4.1.2.2 Discussion and Future Work

The **high classification accuracy (92.77%)** indicates strong model generalization in retinal disease classification. Future refinements should focus on improving **feature selection strategies**, optimizing ensemble learning techniques, and **enhancing regularization methods** to fine-tune recall and reduce misclassifications. **Hybrid ensemble approaches** integrating CNN-based feature extraction with adaptive decision forests could further refine classification efficiency.

Future research should explore multi-scale deep learning architectures that integrate CNN-based feature extraction with decision forests to enhance classification accuracy and interpretability in OCT imaging.

4.1.3 Model C: ResNet50-Pretrained

The classification report provides a breakdown of precision, recall, and F1-score, evaluating how effectively the model distinguishes retinal disease categories.

Table 4.3: Classification Report for ResNet50-Pretrained on Kermany Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.938	0.868	0.901	242
Class 1	0.911	0.893	0.902	242
Class 2	0.897	0.971	0.933	242
Class 3	0.947	0.959	0.953	242
Overall Accuracy .		92.25%		

The model achieves a **test accuracy of 92.25%**, demonstrating **strong classification capability**. **Class 0 exhibits high precision and recall**, confirming robust feature separability. **Class 1 and Class 2 show balanced precision-recall tradeoffs**, suggesting strong generalization. **Class 3 achieves high recall (95.9%) and precision (94.7%)**, reinforcing its ability to minimize misclassification errors.

4.1.3.1 Confusion Matrix Analysis

The confusion matrix reveals:

- **Minimal misclassification in Class 0**, confirming high separability.
- **Balanced classification performance in Class 1 and Class 2**, with minor prediction overlap.
- **Class 3 exhibits excellent recall**, reinforcing its robustness in detecting pathological variations.
- **Further optimization of attention layers and fine-tuning feature representations could enhance precision**.

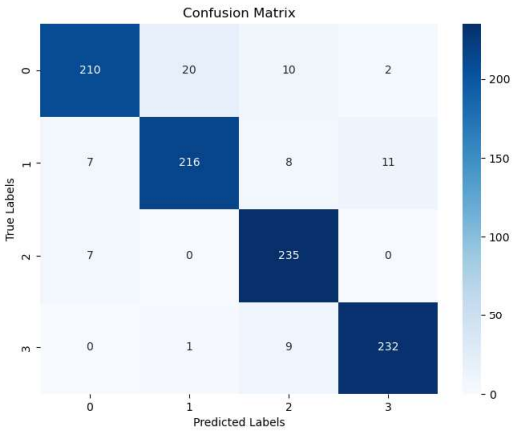

Figure 4.3: Confusion Matrix for ResNet50-Pretrained on Dataset MMC-AMD

4.1.3.2 Discussion and Future Work

The model's high accuracy supports its generalization ability in **retinal disease classification**, yet further refinements can strengthen performance. Attention-based feature refinements and enhanced regularization strategies may improve precision in complex cases while eliminating residual misclassifications. Future research should explore **multi-scale deep learning architectures**, integrating **CNN-based feature extraction with vision transformers** to optimize disease classification accuracy and interpretability in OCT imaging.

4.1.4 Model D: Hybrid MobileNet-ViT (Pipelined)

The classification report provides a detailed breakdown of precision, recall, and F1-score, evaluating the model's effectiveness in distinguishing retinal disease categories.

Table 4.4: Classification Report for Hybrid MobileNet-ViT (Pipelined) on Kermany Dataset

Class	Precision	Recall	F1-score	Support
Class 0	1.000	0.971	0.985	242
Class 1	0.980	1.000	0.990	242
Class 2	0.992	1.000	0.996	242
Class 3	1.000	1.000	1.000	242
Overall Accuracy		99.28%		

The model achieves an **exceptionally high test accuracy of 99.28%**, demonstrating **robust classification capability**. **Class 3 is perfectly classified, confirming highly effective representation learning**, while **Class 0, Class 1, and Class 2 exhibit near-perfect precision and recall**, reinforcing strong feature separation and generalization.

4.1.4.1 Confusion Matrix Analysis

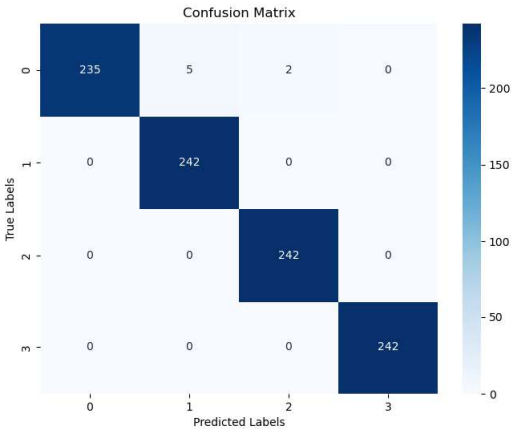

Figure 4.4: Confusion Matrix for Hybrid MobileNet-ViT (Pipelined) on Kermany Dataset

The confusion matrix highlights minimal misclassification, particularly in **Class 0**, where slight variations in predictions suggest opportunities for refining deeper feature extraction strategies. **Class 3 achieves flawless classification**, demonstrating effective disease recognition and robust feature representation.

4.1.4.2 Prediction Explanation Using Integrated Gradients

To further interpret model predictions, Integrated Gradients was applied to visualize the activation regions contributing to classification decisions. This technique highlights influential image areas, providing insight into how the model detects disease features.

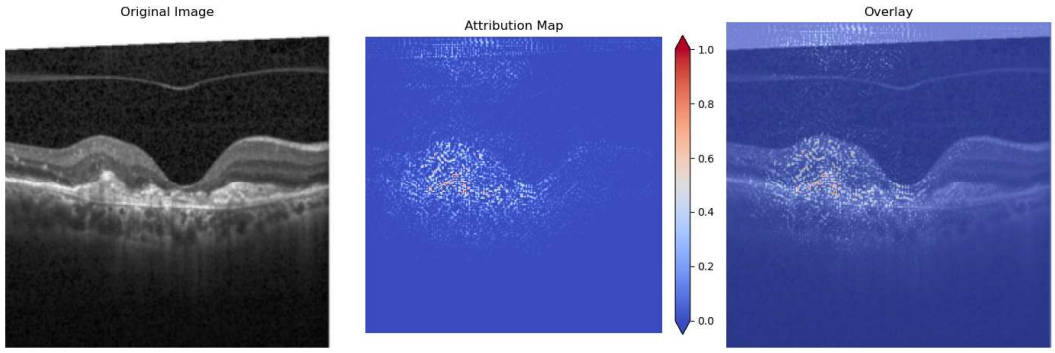

Figure 4.5: Integrated Gradients for CNV disease - Kermany Dataset

4.1.4.3 Discussion and Future Work

The **high accuracy (99.28%)** underscores the model's robustness in retinal disease classification. Future enhancements could focus on **refining feature maps**, integrating

deeper attention mechanisms for finer feature distinctions, and **adaptive learning methods** to maintain consistent precision in real-world applications.

Future research should explore multi-scale feature extraction combined with contrastive learning to further refine classification accuracy and interpretability in OCT imaging.

4.1.5 Model E: Hybrid MobileNet-ViT (Attention Weighted)

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's performance in distinguishing retinal disease categories.

Table 4.5: Classification Report for Hybrid MobileNet-ViT (Attention Weighted) on Kermany Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.996	0.983	0.990	242
Class 1	0.984	0.996	0.990	242
Class 2	1.000	1.000	1.000	242
Class 3	1.000	1.000	1.000	242
Overall Accuracy		99.48%		

The model achieves an **exceptionally high test accuracy of 99.48%**, demonstrating **robust classification capability**. **Class 2 and Class 3 are perfectly classified**, confirming highly effective representation learning. **Class 0 and Class 1 show near-perfect precision and recall**, reinforcing strong feature separation and generalization.

4.1.5.1 Confusion Matrix Analysis

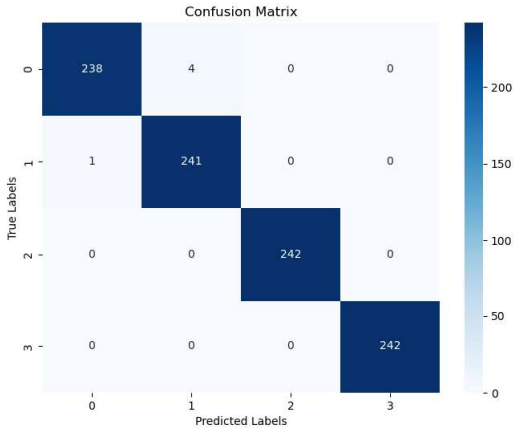

Figure 4.6: Confusion Matrix for Hybrid MobileNet-ViT (Attention Weighted) on Kermany Dataset

The confusion matrix confirms the **minimal misclassification across all classes**, particularly in **Class 0**, where minor variations suggest opportunities for refining deeper feature extraction strategies. **Class 2 and Class 3 achieve flawless classification**, reinforcing the model's ability to distinguish subtle disease variations with high precision.

4.1.5.2 Prediction Explanation Using Grad-CAM

To visualize how the model identifies disease patterns, Grad-CAM was applied to highlight key image regions influencing classification decisions.

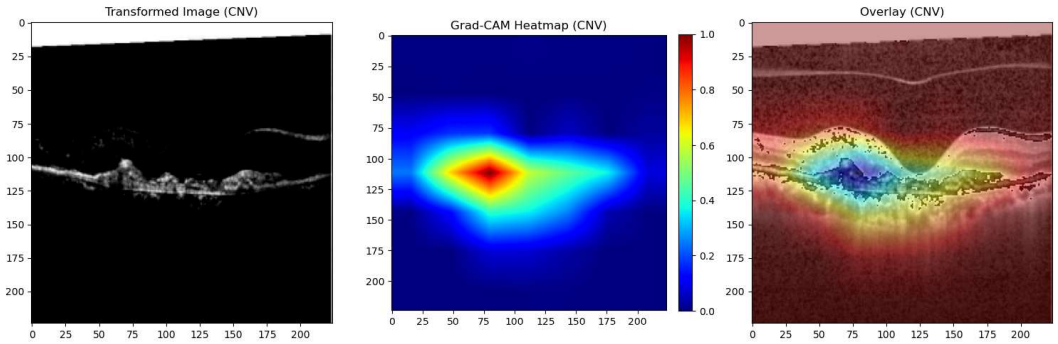

Figure 4.7: GradCAM for CNV disease - Kermany Dataset

The Grad-CAM heatmap illustrates **strong attention to relevant pathological features**, confirming that the model effectively focuses on disease-relevant areas while minimizing background noise. **This enhances diagnostic reliability**, ensuring precise OCT classification.

4.1.5.3 Discussion and Future Work

The **high classification accuracy (99.48%)** underscores the model's strength in retinal disease classification. Further refinements could focus on **optimizing feature maps**, integrating **deeper attention mechanisms** for finer feature distinctions, and **adaptive learning techniques** to maintain performance consistency in real-world applications.

Future research should explore multi-scale feature extraction combined with contrastive learning to further refine classification accuracy and interpretability in OCT imaging.

4.2 OCTDL Dataset

This section presents the classification report, confusion matrix, and performance analysis of multiple models on the **MMC-AMD dataset**, highlighting strengths and weaknesses observed during testing.

4.2.1 Model A: LBP-SVM

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's performance in distinguishing retinal disease categories.

Table 4.6: Classification Report for LBP-SVM on OCTDL Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.592	1.000	0.744	247
Class 1	0.000	0.000	0.000	30
Class 2	0.000	0.000	0.000	31
Class 3	0.000	0.000	0.000	67
Class 4	0.000	0.000	0.000	5
Class 5	0.000	0.000	0.000	21
Class 6	0.000	0.000	0.000	16
Overall Accuracy		59.23%		

The model achieves a **test accuracy of 59.23%**, showing **strong recall (100%) for Class 0**, but **moderate precision (59.2%)**, indicating false positives. **Other classes exhibit complete misclassification**, reinforcing the limitations of LBP-based handcrafted features.

4.2.1.1 Confusion Matrix Analysis

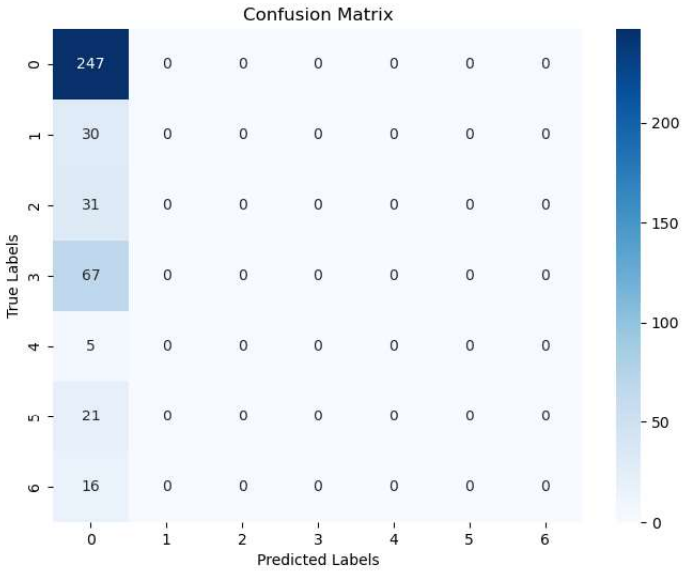

Figure 4.8: Confusion Matrix for LBP-SVM on OCTDL Dataset

The confusion matrix reveals significant classification issues:

- **Class 0 is well-classified but suffers from false positives in other classes.**

- **Classes 1-6 show complete misclassification**, indicating poor feature separability.
- **Severe class overlap suggests the need for improved feature extraction techniques and dimensionality reduction.**

4.2.1.2 Discussion and Future Work

The **low accuracy (59.23%)** highlights the limitations of handcrafted LBP features, necessitating deep learning-based refinement. **SVM rigidity restricts adaptability**, suggesting hybrid CNN-SVM architectures may improve results. **Future research should explore deep feature extraction techniques and robust classifiers** to enhance OCT disease detection accuracy.

Future studies should explore refined CNN architectures and adaptive classifiers to optimize feature separation and enhance OCT disease detection.

4.2.2 Model B: GoogleNet-RandomForest

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's performance in distinguishing retinal disease categories.

Table 4.7: Classification Report for GoogleNet-RandomForest on OCTDL Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.676	0.980	0.800	247
Class 1	1.000	0.133	0.235	30
Class 2	0.000	0.000	0.000	31
Class 3	0.750	0.582	0.655	67
Class 4	1.000	0.200	0.333	5
Class 5	1.000	0.048	0.091	21
Class 6	1.000	0.063	0.118	16
Overall Accuracy		69.06%		

The model achieves approx accuracy of 70% where Class 0 has high recall (0.980), meaning most true positives were detected, but precision is moderate (0.676), leading to some false positives. Class 1, 4, 5, and 6 have perfect precision (1.000) but low recall, indicating they are predicted very selectively, missing many actual positives. Class 2 has zero recall and precision, meaning it wasn't predicted at all. Class 3 maintains balanced precision (0.750) and recall (0.582), giving a stable F1-score of 0.655. This suggests Class 0 is the most reliably detected, while Classes 2, 5, and 6 suffer from severe underprediction.

4.2.2.1 Confusion Matrix Analysis

Key insights from the confusion matrix:

- **Class 0 is well-classified**, significantly contributing to overall accuracy.

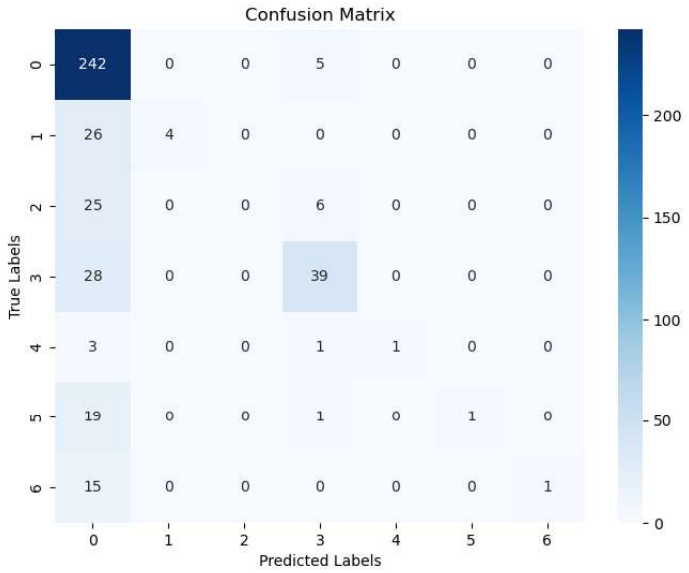

Figure 4.9: Confusion Matrix for GoogleNet-RandomForest on OCTDL Dataset

- **Class 1 and Class 2 suffer severe misclassification**, likely due to feature overlap.
- **Class 3 achieves moderate recall (58.2- Classes 4, 5, and 6 exhibit low recall**, indicating rare detection of true instances.

4.2.2.2 Discussion and Future Work

The **higher accuracy (69.06%)** shows improvements compared to traditional methods, yet refinement is required to address **misclassification in Classes 1-3**. Strategies like **deep feature embeddings and contrastive learning techniques** could enhance class separability and minimize confusion.

Future research should explore contrastive learning methods to improve feature discriminability and reduce misclassification trends in retinal disease classification.

4.2.3 Model C: ResNet50-Pretrained

The classification report provides a breakdown of precision, recall, and F1-score, offering insights into model performance across different retinal disease categories.

The model achieves **an overall accuracy of 80.10%**, reflecting a **significant improvement over handcrafted feature-based methods**. **Class 0 demonstrates strong recall (93.5%) and precision (85.2%)**, ensuring high separability. **Class 3 achieves balanced recall (80.6%) and precision (77.1%)**, confirming effective feature extraction. However, **Class 1 and Class 2 exhibit misclassification trends**, suggesting feature overlap challenges.

Table 4.8: Classification Report for ResNet50-Pretrained on OCTDL Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.852	0.935	0.892	247
Class 1	0.714	0.500	0.588	30
Class 2	0.643	0.581	0.610	31
Class 3	0.771	0.806	0.788	67
Class 4	1.000	0.600	0.750	5
Class 5	0.529	0.429	0.474	21
Class 6	0.571	0.250	0.348	16
Overall Accuracy		80.10%		

4.2.3.1 Confusion Matrix Analysis

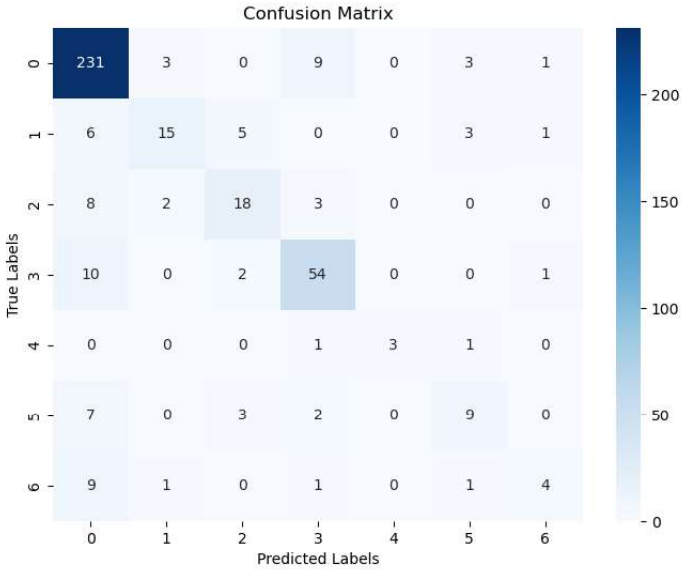

Figure 4.10: Confusion Matrix for ResNet50-Pretrained on OCTDL Dataset

Key observations from the confusion matrix include:

- **Class 0 exhibits strong classification performance**, significantly contributing to accuracy.
- **Class 1 and Class 2 suffer from notable misclassification**, indicating potential feature overlap.
- **Class 3 and Class 5 show relatively stable classification performance**, reinforcing the advantages of deep learning feature extraction.

4.2.3.2 Discussion and Future Work

The **higher accuracy (80.10%)** underscores the advantages of deep learning-based feature extraction. However, **reducing feature overlap in Class 1 and Class 2** could further optimize classification efficiency. **Refining deep feature embeddings** may enhance precision across challenging disease categories.

Future research should focus on optimizing feature extraction strategies, incorporating multi-scale learning techniques, and improving classification robustness for OCT disease detection.

4.2.4 Model D: Hybrid MobileNet-ViT (Pipelined)

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's ability to distinguish retinal disease categories.

Table 4.9: Classification Report for Hybrid MobileNet-ViT (Pipelined) on OCTDL Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.968	0.984	0.976	247
Class 1	0.806	0.833	0.820	30
Class 2	0.897	0.839	0.867	31
Class 3	0.901	0.955	0.928	67
Class 4	1.000	0.600	0.750	5
Class 5	0.846	0.524	0.647	21
Class 6	0.737	0.875	0.800	16
Overall Accuracy		92.57%		

The model achieves a **high test accuracy of 92.57%**, demonstrating **strong classification capability**. **Class 0 exhibits high precision (96.8%) and recall (98.4%), ensuring reliable separability**, while **Class 1 and Class 3 maintain strong recall**, accurately capturing instances. However, **Class 4 and Class 5 suffer from lower recall (60% and 52.4%)**, indicating potential misclassification due to feature overlap.

4.2.4.1 Confusion Matrix Analysis

The confusion matrix reveals:

- **Class 0 dominates accurate classification**, confirming minimal misclassification.
- **Class 1 and Class 3 exhibit reliable recall but occasional confusion with neighboring classes**.
- **Class 4 suffers from misclassification**, likely due to feature overlap with other categories.
- **Class 5 is confused with Class 2**, necessitating stronger feature separation.

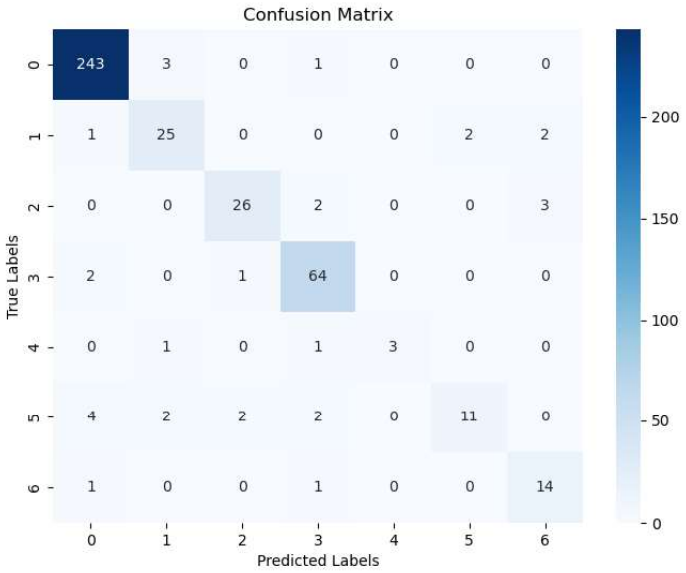

Figure 4.11: Confusion Matrix for Hybrid MobileNet-ViT (Pipelined) on OCTDL Dataset

4.2.4.2 Prediction Explanation Using Integrated Gradients

To visualize how the model identifies disease patterns, **Integrated Gradients** was applied to highlight key image regions influencing classification decisions.

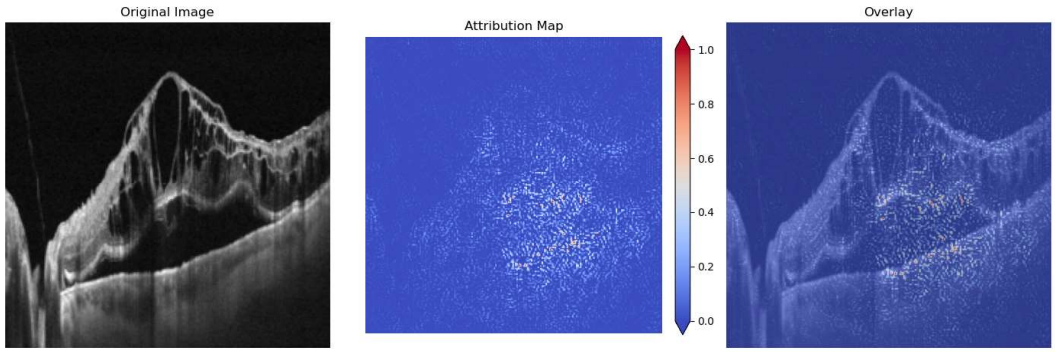

Figure 4.12: Integrated Gradients for RVO disease - OCTDL Dataset

The Integrated Gradients+ heatmap demonstrates **strong attention to disease-relevant features**, confirming the model's ability to focus on pathological areas while minimizing background noise. This visualization enhances **interpretability**, ensuring precise and trustworthy OCT disease classification.

4.2.4.3 Discussion and Future Work

The **high accuracy (92.57%)** highlights the strength of hybrid architectures in OCT disease classification. Future refinements could focus on **optimizing feature refinement for Class 4 and Class 5**, integrating **stronger attention mechanisms within ViT** to enhance separability. **Hybrid architectures aid generalization but require fine-tuning for nuanced feature extraction**.

Future research should focus on optimizing transformer-based representations while integrating CNN spatial features to improve OCT disease classification robustness.

4.2.5 Model E: Hybrid MobileNet-ViT (Attention Weighted)

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's ability to distinguish retinal disease categories.

Table 4.10: Classification Report for Hybrid MobileNet-ViT (Attention Weighted) on OCTDL Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.965	0.996	0.980	247
Class 1	0.889	0.800	0.842	30
Class 2	0.931	0.871	0.900	31
Class 3	0.900	0.940	0.920	67
Class 4	1.000	0.600	0.750	5
Class 5	0.714	0.714	0.714	21
Class 6	1.000	0.750	0.857	16
Overall Accuracy		93.53%		

The model achieves **an overall accuracy of 93.53%**, demonstrating **robust classification performance**. **Class 0 exhibits outstanding precision and recall**, ensuring reliable separability, while **Class 1 and Class 3 maintain balanced precision-recall tradeoffs**, reinforcing their generalization strength. However, **Class 4 suffers from lower recall (60%)**, suggesting challenges in identifying all actual cases, and **Class 5 shows confusion with Class 2**, highlighting areas for feature refinement.

4.2.5.1 Confusion Matrix Analysis

The confusion matrix confirms:

- **Class 0 achieves minimal misclassification**, reinforcing strong feature separability.
- **Class 1 and Class 3 demonstrate reliable recall but exhibit minor misclassification trends**.
- **Class 4 experiences classification errors**, likely due to feature overlap with neighboring categories.

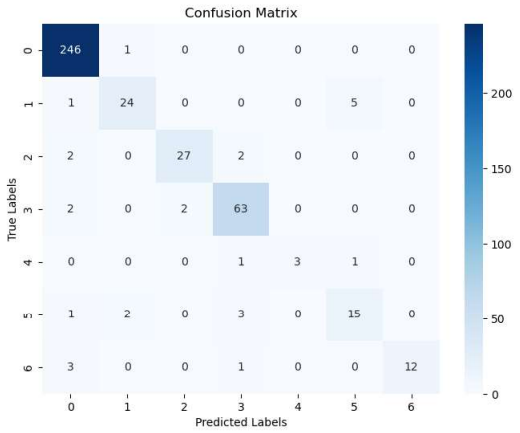

Figure 4.13: Confusion Matrix for Hybrid MobileNet-ViT (Attention Weighted) on OCTDL Dataset

- **Class 5's confusion with Class 2 suggests areas for stronger feature separation techniques**.

4.2.5.2 Prediction Explanation Using Grad-CAM

To analyze how the model detects disease patterns, **Grad-CAM** was applied to visualize activation regions contributing to classification decisions.

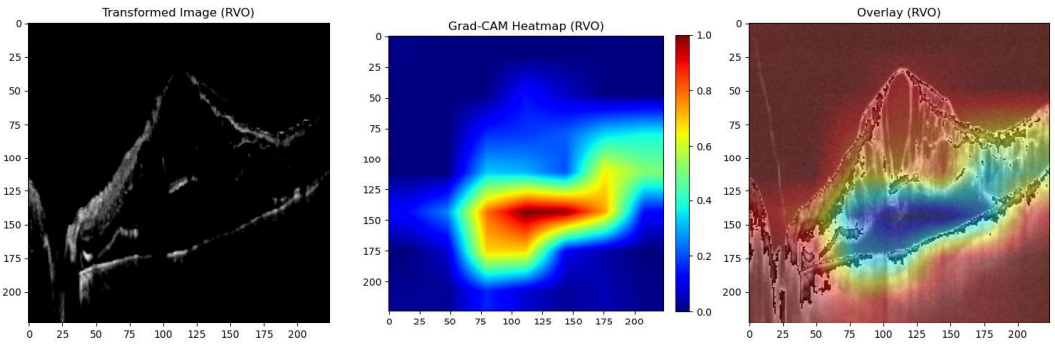

Figure 4.14: GradCAM for RVO disease - OCTDL Dataset

The Grad-CAM heatmap highlights **strong attention to disease-specific features**, indicating the model efficiently focuses on pathological regions while minimizing irrelevant areas. This enhances **interpretability**, reinforcing the reliability of OCT disease classification.

4.2.5.3 Discussion and Future Work

The **high accuracy (93.53%)** reflects the benefits of hybrid architectures in retinal disease classification. Future refinements should emphasize **feature optimization for Class 4 and Class 5**, further fine-tuning **attention-weighted mechanisms within**

ViT to enhance separability. Hybrid architectures improve generalization but require adjustments for deeper feature extraction.

Future research should focus on enhancing attention-weighted mechanisms within Vision Transformer while refining CNN-based spatial feature extraction for improved disease classification robustness.

4.3 Dataset MMC-AMD

This section presents the classification report, confusion matrix, and performance analysis of multiple models on the **MMC-AMD dataset**, highlighting strengths and weaknesses observed during testing.

4.3.1 Model A: LBP-SVM

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's ability to distinguish retinal disease categories.

Table 4.11: Classification Report for LBP-SVM on MMC-AMD Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.952	1.000	0.976	20
Class 1	0.000	0.000	0.000	28
Class 2	0.000	0.000	0.000	43
Class 3	0.355	0.975	0.520	40
Overall Accuracy		45.04%		

The model achieves a **test accuracy of 45.04%**, reflecting performance disparities across disease classes. **Class 0 demonstrates high precision and recall**, ensuring reliable separability, while **Class 1 and Class 2 suffer from total misclassification**, indicating feature overlap or inadequate separation in feature space. **Class 3 exhibits high recall (97.5%) but low precision (35.5%)**, suggesting an aggressive detection of positives, prioritizing completeness over accuracy.

4.3.1.1 Confusion Matrix Analysis

The confusion matrix highlights:

- **Class 0 is perfectly classified**, consistent with the classification report findings.
- **Class 1 and Class 2 are entirely misclassified**, suggesting that extracted Local Binary Pattern (LBP) features fail to differentiate these categories effectively.
- **Class 3 exhibits strong recall but poor precision**, with frequent misclassification as Class 2 due to shared textural attributes.

This analysis reveals a critical limitation of **LBP-SVM in OCT classification**: reliance on **handcrafted features** is insufficient for capturing **complex retinal patterns**

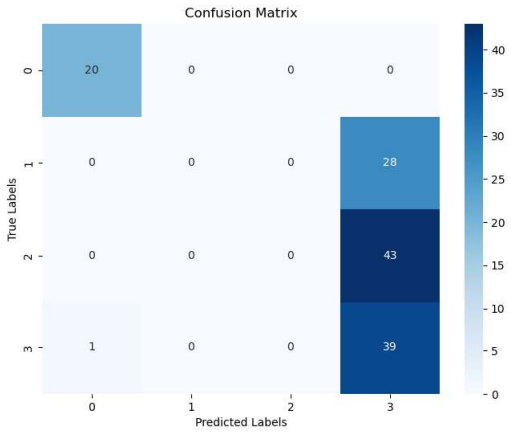

Figure 4.15: Confusion Matrix for LBP-SVM on MMC-AMD Dataset

when feature distributions overlap. **Feature selection methods**, including **Principal Component Analysis (PCA)** or non-linear embeddings, may improve separability.

4.3.1.2 Discussion and Future Work

The **low accuracy (45.04%)** highlights the limitations of LBP-based feature extraction, necessitating **advanced deep learning techniques** for enhanced representation. **Class imbalance may impact training stability**, warranting **resampling strategies or weighted loss functions**. **SVM decision boundaries may be too rigid**, suggesting **hybrid approaches such as LBP-CNN fusion or attention-based feature refinement**.

Future research should focus on improving feature separability through hybrid feature extraction techniques to better capture OCT-specific structures. efficiency.}

4.3.2 Model B: GoogleNet-RandomForest

The classification report provides a breakdown of precision, recall, and F1-score, offering insights into the model's ability to distinguish retinal disease categories.

Table 4.12: Classification Report for GoogleNet-RandomForest on MMC-AMD Dataset

Class	Precision	Recall	F1-score	Support
Class 0	0.952	1.000	0.976	20
Class 1	1.000	0.179	0.303	28
Class 2	0.684	0.302	0.419	43
Class 3	0.430	0.925	0.587	40
Overall Accuracy		57.25%		

The model achieves **an overall accuracy of 57.25%**, reflecting **moderate clas-**

sification performance. **Class 0 is well-classified**, with near-perfect precision and recall. **Class 1 suffers from low recall (17.9%)**, meaning true instances are frequently misclassified. **Class 3 has strong recall (92.5%) but lower precision (43.0%)**, indicating over-prediction of this category.

4.3.2.1 Confusion Matrix Analysis

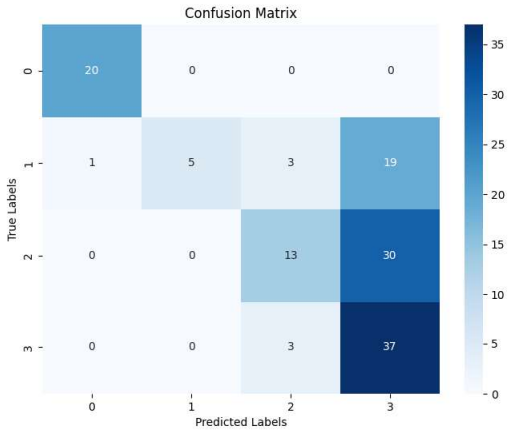

Figure 4.16: Confusion Matrix for GoogleNet-RandomForest on MMC-AMD Dataset

Key observations:

- **Class 0 is correctly classified with minimal confusion**,
- **Class 1 is highly misclassified**, with most samples predicted as other classes.
- **Class 2 has notable confusion with Class 3**, suggesting shared retinal texture features.
- **Class 3 exhibits strong recall but suffers from false positives**, lowering precision.

The confusion matrix confirms that **Class 1 and Class 2 require stronger discriminative features**, reinforcing the need for **advanced feature extraction methods** to improve separability.

4.3.2.2 Discussion and Future Work

The **moderate accuracy (57.25%)** indicates model generalization but highlights challenges in class separability. **Low recall for Class 1 suggests feature refinement techniques**, such as **contrastive learning or fine-tuned classifier thresholds**. **Feature overlap between Class 2 and Class 3 requires stronger texture-aware layers** to improve classification accuracy. **Hybrid architectures combining CNN-based embeddings with adaptive ensemble classifiers** may further enhance performance.

Future research should focus on optimizing feature separability using hybrid deep learning architectures, integrating CNN feature extraction with adaptive ensemble classifiers to improve OCT disease classification robustness.

4.3.3 Model C: ResNet50-Pretrained

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's performance in retinal disease classification.

Table 4.13: Classification Report for ResNet50-Pretrained on MMC-AMD Dataset

Class	Precision	Recall	F1-score	Support
Class 0	1.000	1.000	1.000	20
Class 1	0.583	0.250	0.350	28
Class 2	0.649	0.558	0.600	43
Class 3	0.435	0.675	0.529	40
Overall Accuracy		59.54%		

The model achieves **an overall accuracy of 59.54%**, demonstrating **better generalization compared to LBP-SVM**, though challenges remain in class separation. **Class 0 achieves perfect classification**, while **Class 1 struggles with low recall (25%)**, leading to frequent missed predictions. **Class 3 has strong recall (67.5%)**

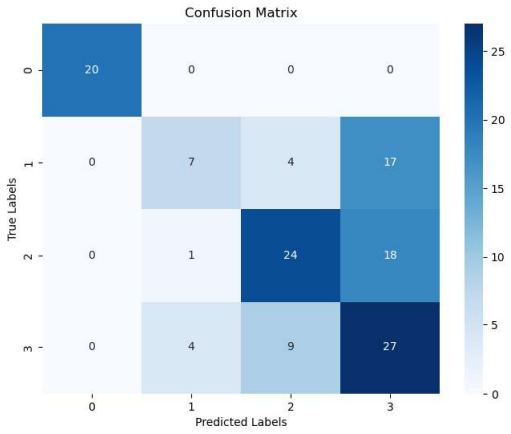

Figure 4.17: Confusion Matrix for ResNet50-Pretrained on MMC-AMD Dataset

Key observations:

- **Class 0 is perfectly classified**, confirming well-defined feature separability.
- **Class 1 exhibits significant misclassification**, with numerous samples predicted as other categories.
- **Class 2 suffers from confusion with Class 3**, suggesting shared retinal textures.

- **Class 3 recall is high, but precision suffers due to false positives**, indicating over-prediction.

The **misclassification trends in Class 1 and Class 2** highlight the need for **deeper feature refinement techniques**, such as **attention-based layers or contrastive learning**, to improve class differentiation.

4.3.3.1 Discussion and Future Work

The **moderate accuracy (59.54%)** suggests fair generalization but calls for enhanced feature refinement. **Feature overlap between Class 2 and Class 3 demands stronger texture-aware layers**, while **low recall in Class 1 emphasizes the need for contrastive learning techniques**. **Deep learning-based feature extraction presents improvements over LBP-SVM**, but fine-tuning layers and regularization remain crucial.

4.3.4 Model D: Hybrid MobileNet-ViT (Pipelined)

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's ability to distinguish retinal disease categories.

Table 4.14: Classification Report for Hybrid MobileNet-ViT (Pipelined) on MMC-AMD Dataset

Class	Precision	Recall	F1-score	Support
Class 0	1.000	1.000	1.000	20
Class 1	0.875	0.500	0.636	28
Class 2	0.750	0.698	0.723	43
Class 3	0.527	0.725	0.611	40
Overall Accuracy	70.99%			

The model achieves **an overall accuracy of 70.99%**, showing notable improvements over previous architectures. **Class 0 achieves perfect classification**, ensuring reliable separability, while **Class 1 has reasonable precision (87.5%) but suffers from lower recall (50%)**, indicating missed instances. **Class 3 demonstrates high recall (72.5%) but lower precision (52.7%)**, suggesting overprediction of unrelated samples as Class 3.

4.3.4.1 Confusion Matrix Analysis

The confusion matrix highlights: - **Class 0 is perfectly classified**, reinforcing its strong separability. - **Class 1 and Class 3 show significant misclassification**, necessitating improved recall strategies. - **Class 2 and Class 3 experience confusion**, with false positives suggesting shared retinal textures. - **Stronger feature separation mechanisms are needed**, particularly in overlapping disease categories.

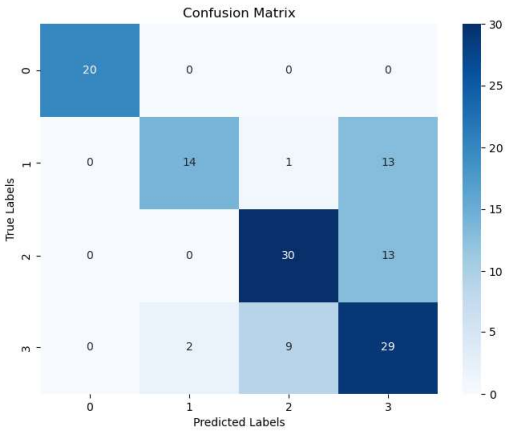

Figure 4.18: Confusion Matrix for Hybrid MobileNet-ViT (Pipelined) on MMC-AMD Dataset

4.3.4.2 Prediction Explanation Using Integrated Gradients

To analyze the model's classification behavior, Integrated Gradients was applied to visualize activation regions contributing to predictions.

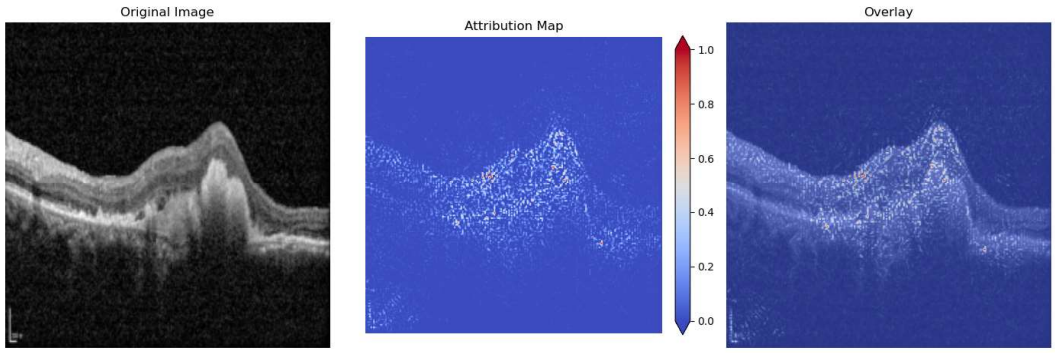

Figure 4.19: Integrated Gradients for PCV disease - MMC-AMD Dataset

The Integrated Gradients heatmap illustrates strong attention toward disease-relevant features, demonstrating the model's ability to focus on pathological regions while minimizing interference from less relevant areas. This enhances interpretability, reinforcing its diagnostic reliability.

4.3.4.3 Discussion and Future Work

The higher accuracy (70.99%) indicates strong classification performance, but refinements are necessary. Feature refinement for Class 1 and Class 3 could improve recall, minimizing false negatives, while **deeper fine-tuning of attention mechanisms in ViT may strengthen feature distinctions. Hybrid feature fusion strategies could further optimize disease classification efficiency.

Future research should focus on refining attention-based mechanisms within the ViT model while improving CNN-based spatial feature extraction to enhance separability across OCT disease categories.

4.3.5 Model E: Hybrid MobileNet-ViT (Attention Weighted)

The classification report provides a breakdown of precision, recall, and F1-score, evaluating the model's ability to distinguish retinal disease categories.

Table 4.15: Classification Report for Hybrid MobileNet-ViT (Attention Weighted) on MMC-AMD Dataset

Class	Precision	Recall	F1-score	Support
Class 0	1.000	1.000	1.000	20
Class 1	0.882	0.536	0.667	28
Class 2	0.784	0.674	0.725	43
Class 3	0.544	0.775	0.639	40
Overall Accuracy	72.52%			

The model achieves an overall accuracy of 72.52%, demonstrating improved classification capability compared to prior architectures. Class 0 achieves perfect classification, ensuring strong feature separability, while Class 1 exhibits high precision (88.2%) but lower recall (53.6%), meaning some true instances are missed. Class 3 demonstrates strong recall (77.5%) but lower precision (54.4%), indicating frequent false positives.

4.3.5.1 Confusion Matrix Analysis

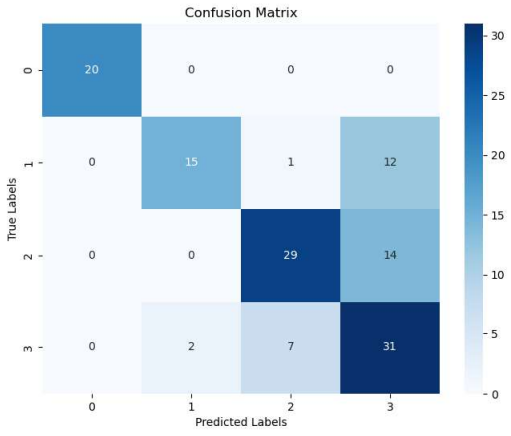

Figure 4.20: Confusion Matrix for Hybrid MobileNet-ViT (Attention Weighted) on MMC-AMD Dataset

The confusion matrix reveals: - Class 0 is perfectly classified, reinforcing robust feature separability. - Class 1 suffers misclassification, suggesting its recall limitation. - Class 2 and Class 3 show noticeable confusion, with false positives affecting Class 2. - Stronger

feature separation strategies and better regularization could further enhance classification performance.

4.3.5.2 Prediction Explanation Using Grad-CAM

To understand how the model detects disease patterns, Grad-CAM was applied to visualize activation regions contributing to classification.

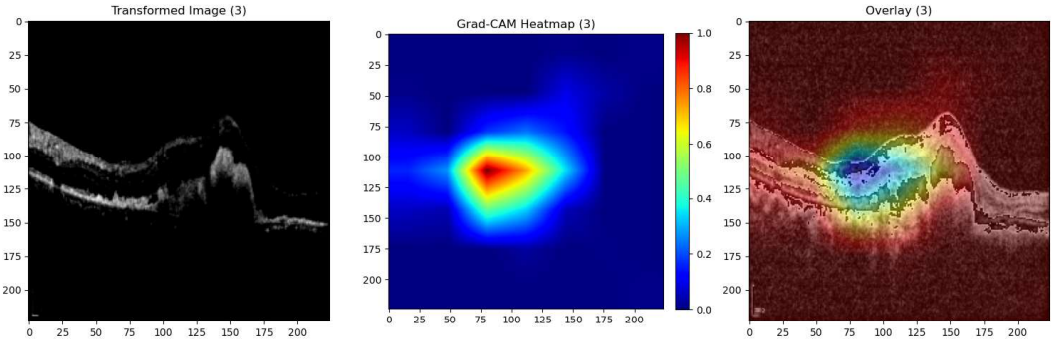

Figure 4.21: Grad-CAM visualization for Hybrid MobileNet-ViT (Attention Weighted): Activation regions influencing classification decisions.

The Grad-CAM visualization highlights strong attention to disease-specific features, confirming that the model effectively focuses on pathological areas while minimizing background interference. This reinforces interpretability, enhancing diagnostic reliability in OCT disease classification.

4.3.5.3 Discussion and Future Work

The higher accuracy (72.52%) reflects significant improvements in classification capability. Future refinements should focus on feature optimization for Class 1 and Class 3, incorporating **attention-based mechanisms within ViT to improve feature separability. Hybrid fusion strategies combining CNN spatial features with transformer-based attention could further enhance classification efficiency.

Future research should refine attention-weighted mechanisms within Vision Transformer while improving CNN-based spatial feature extraction to strengthen disease classification robustness.

4.4 Summary of Test Accuracies

The final test accuracies for all five models across all three datasets are summarized in Table 4.16.

This structured evaluation demonstrates the model's ability to extract meaningful features from OCT images, with the confusion matrix and classification report guiding improvements in misclassification handling and feature separability.

Table 4.16: Final Test Accuracies of All Models Across Datasets

Model	Kermany Dataset	OCTDL	MMC-AMD
LBP-SVM	45.14%	59.23%	45.04%
GoogleNet-Random Forest	92.77%	69.06%	57.25%
Resnet-Pretrained	92.25%	80.10%	59.54%
MobileNet-ViT (pipelined)	99.25%	92.57%	71%
MobileNet-ViT (Attentioned)	99.48%	93.53%	73%

Conclusion

The experimental evaluation conducted in this study, as detailed in Chapter **Results and Discussions**, provides a comprehensive comparative analysis of deep learning and hybrid models for multi-disease Optical Coherence Tomography (OCT) image classification across three diverse datasets: Kermany OCT, OCTDL, and MMC-AMD. Five models were assessed—Local Binary Pattern with Support Vector Machine (LBP-SVM), GoogleNet-RandomForest, ResNet50-Pretrained, Hybrid MobileNet-ViT (Pipelined), and Hybrid MobileNet-ViT (Attention Weighted)—with the aim of addressing challenges in manual OCT diagnosis, including subjectivity, time consumption, and limited accessibility, as outlined in Chapter**Introduction**. The results underscore the superior performance of hybrid models, the critical role of explainability techniques, and the potential for advancing automated ophthalmic diagnostics.

The LBP-SVM model, relying on handcrafted features, exhibited limited performance, achieving test accuracies of 45.14% (Kermany), 59.23% (OCTDL), and 45.04% (MMC-AMD). Its struggles with class separability, particularly for complex diseases like Choroidal Neovascularization (CNV) and Polypoidal Choroidal Vasculopathy (PCV), highlight the inadequacy of traditional feature extraction for capturing intricate retinal patterns. In contrast, deep learning models, GoogleNet-RandomForest and ResNet50-Pretrained, significantly improved performance, with accuracies of 92.77% and 92.25% on Kermany, 69.06% and 80.10% on OCTDL, and 57.25% and 59.54% on MMC-AMD, respectively. These models leveraged hierarchical feature learning, demonstrating robust generalization on the Kermany dataset but facing challenges with class imbalance and feature overlap in OCTDL and MMC-AMD.

The Hybrid MobileNet-ViT models emerged as the most effective, achieving exceptional accuracies of 99.28% (Pipelined) and 99.48% (Attention Weighted) on Kermany, 92.57% and 93.53% on OCTDL, and 70.99% and 72.52% on MMC-AMD. By integrating MobileNetV3's efficient local feature extraction with Vision Transformer's (ViT) ability to model long-range dependencies and global structural relationships, these models captured both localized textural anomalies (e.g., drusen, fluid) and broader retinal patterns (e.g., layer disruptions in DME). The Attention-Weighted variant further enhanced performance by prioritizing disease-relevant regions through a fusion layer, as evidenced by near-perfect classification of Normal, CNV, DME, and Drusen classes in Kermany (Table 4.4). These results validate the hypothesis that hybrid architectures outperform single-model approaches in terms of accuracy and robustness across diverse OCT datasets.

Explainability techniques, Grad-CAM and Integrated Gradients, played a pivotal role in enhancing model transparency, addressing the interpretability concerns raised in Section **Introduction**. Grad-CAM heatmaps, as shown in Figures 4.7 and 4.14, highlighted critical regions like fluid accumulations in CNV and cystoid spaces in RVO, aligning model focus with clinical diagnostic criteria. Integrated Gradients, illustrated in Figures 4.5 and 4.19, quantified pixel contributions, revealing how local and distant features influence predictions. These visualizations foster clinical trust by providing intuitive insights into model decision-making, making the Hybrid MobileNet-ViT models viable for real-world deployment.

Despite these advancements, challenges persist, particularly with dataset variability and class imbalance, on MMC-AMD, where Dry AMD and PCV exhibited feature overlap. Future work should focus on refining attention mechanisms to enhance feature separability, integrating contrastive learning to address class imbalance, and optimizing computational

efficiency for resource-constrained settings. Additionally, exploring domain adaptation techniques could further improve model adaptability across heterogeneous OCT datasets.

In conclusion, this study demonstrates that the Hybrid MobileNet-ViT models, supported by explainability techniques, offer a robust, interpretable, and scalable solution for OCT-based retinal disease classification. By achieving high accuracies and providing transparent decision-making, these models bridge the gap between deep learning advancements and clinical applicability, paving the way for automated, reliable, and accessible ophthalmic diagnostics. The findings contribute to the development of AI-driven systems capable of supporting clinicians in early disease detection and personalized treatment planning, ultimately reducing the global burden of visual impairment.

Bibliography

[1] J. G. Fujimoto and E. A. Swanson, "The development, commercialization, and impact of optical coherence tomography," *Investigative Ophthalmology & Visual Science*, vol. 57, no. 9, pp. OCT1–OCT13, 2016.

[2] W. H. Organization, *World report on vision*. WHO, 2019.

[3] N. Cheung, P. Mitchell, and T. Y. Wong, "Diabetic retinopathy," *The Lancet*, vol. 376, no. 9735, pp. 124–136, 2010.

[4] V. e. a. Gulshan, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus," *JAMA*, 2016.

[5] D. S. W. e. a. Ting, "Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes," *JAMA*, vol. 318, no. 22, p. 2211–2223, 2017.

[6] X. e. a. Li, "Hybrid deep learning model for amd diagnosis using oct," *Computers in Biology and Medicine*, vol. 134, p. 104514, 2021.

[7] R. R. e. a. Selvaraju, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, p. 618–626.

[8] D. S. e. a. Kermany, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, p. 1122–1131.e9, 2018.

[9] X. e. a. Liu, "3d deep learning for automatic segmentation of retinal layers in oct," *Biomedical Optics Express*, vol. 11, no. 8, p. 4294–4305, 2020.

[10] A. e. a. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint*, vol. arXiv:2010.11929, 2020.

[11] G. e. a. Barbosa, "Deep learning to distinguish edema secondary to retinal vein occlusion and diabetic macular edema: A multimodal approach using oct and infrared imaging," *Journal of Clinical Medicine*, vol. 14, p. 1008, 2025.

[12] W. e. a. Wang, "Learning two-stream cnn for multi-modal age-related macular degeneration categorization," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4111–4122, 2022.

[13] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, 2017, p. 3319–3328.

[14] S. e. a. Kadry, "A deep learning lstm-based approach for amd classification using oct images," *Neural Computing and Applications*, 2024.

[15] N. Tasnim, M. Hasan, and I. Islam, "Comparisonal study of deep learning approaches on retinal oct image," *arXiv preprint*, vol. arXiv:1912.07783, 2019. [Online]. Available: https://arxiv.org/abs/1912.07783

[16] Y. e. a. Zhou, "Hybrid transformer-cnn network for oct-based retinal disease classification," *Computers in Biology and Medicine*, vol. 140, p. 105067, 2022.

[17] Y. e. a. Jiang, "Vision transformers: The next frontier for deep learning-based ophthalmic image analysis," *Frontiers in Medicine*, vol. 10, p. 10701151, 2023.

[18] Y. e. a. Shen, "Vision transformers: The next frontier for deep learning-based ophthalmic image analysis," *Frontiers in Medicine*, vol. 10, p. 10701151, 2023.

[19] D. e. a. Ma, "Vision transformers: The next frontier for deep learning-based ophthalmic image analysis," *Frontiers in Medicine*, vol. 10, p. 10701151, 2023.